***Compare partitions***

SPSS macros by Kirill Orlov

kior@akado.ru, ttnphns@gmail.com

<https://www.spsstools.net/en/KO-spssmacros>

All rights reserved

*Сравнение классификаций/кластеризаций.* Вычисление разных мер подобия группирований: внешние кластерные критерии, индексы правильности и согласия классификаций. Идентификация одинаковых или схожих групп между группированиями.

*Прочтите «*[*О SPSS макросах*](https://www.spsstools.net/ru/KO-aboutmacros)*» что они такое и как их запускать.*

*Ошибка “Protected directory”.* Некоторые из макросов, описанных в текущем документе, пишут временные файлы на жесткий диск. Если вы не обладаете полными правами Администратора вашего компьютера, это может вызвать ошибку, сообщающую среди прочего: *“SPSS Statistics cannot access a file... specifies a protected directory...”* и значащую, что дефолтная директория, какую макрос хочет использовать, защищена на вашем ПК. Чтобы решить эту проблему, в окне синтаксиса скомандуйте: CD 'myfolder'., где 'myfolder' есть путь/имя некоторой папки, куда вам разрешено сохранять файлы.

Пусть вы хотите получить показатель согласия (подобия) между двумя разбиениями *U* и *V* одного и того же набора объектов на (непересекающиеся) группы. Разбиения *U* и *V* похожи тогда, когда в обоих разбиениях есть одни и те же по составу объектов группы. Перед сравнением разбиений возникает вопрос: ***мы знаем заранее, какой группе из U отвечает именно какая группа из V, или не знаем?*** Очевидно, что если и *U* и *V* это помеченные (labeled, идентифицированные) классы, то (1) мы знаем заранее, какому одному классу из *U* соответствует какой один класс из *V*. Но если *U* и *V* это непомеченные (unlabeled) кластеры либо одно из двух разбиений это классы, а другое это кластеры, то возникает вопрос, нужно ли нам предварительно (2) находить такие взаимно-однозначные соответствия групп между разбиениями (3) или же можно обойтись без явного взаимно-однозначного спаривания групп?

* Варианту (3) отвечает макрос [!KO\_CLUAGREE](#_МАКРОС_!PROXQNT:_РАЗНЫЕ) для вычисления мер подобия между «непомеченными» разбиениями (т.е. взаимосоответствие групп между разбиениями не установлено). Этот макрос вычисляет меры подобия, которые либо не опираются на соответствие групп между разбиениями *U* и *V*, либо соответствие между их группами как-то определяется имплицитно самой формулой меры подобия. Типичный пример использования – сравнение кластеризаций или кластеризации с классификацией.
* Варианту (1) отвечает макрос [!KO\_CLASAGREE](#_МАКРОС_!CLASAGREE:_МЕРЫ) для вычисления мер подобия между «помеченными» разбиениями (т.е. взаимосоответствие групп между разбиениями установлено) . Этот макрос вычисляет меры подобия, которые базируются на информации, «правильно» или «неправильно» классифицирован объект в разбиении *V* относительно разбиения *U*. Типичный пример использования – сравнение классификаций.
* Варианту (2) отвечает макрос [!KO\_GRMATCH](#_МАКРОС_!GRMATCH:_ИДЕНТИФИКАЦИЯ_1) для установления взаимосоответствующих групп между разбиениями. Если надо, вы можете после перекодировки групп этим макросом использовать макрос !KO\_CLASAGREE.

# МАКРОС !KO\_CLUAGREE: МЕРЫ СРАВНЕНИЯ РАЗБИЕНИЙ-«КЛАСТЕРИЗАЦИЙ»

Version 1, Mar 2019. Tested on SPSS Statistics 17, 20, 22.

!KO\_cluagree vars= v1 v2 v3 v4 /\*Переменные, являющие собой группирования наблюдений;

/\*слово WITH возможно после первого имени

/measure= DICE /\*Укажите меру(ы): DICE, OCHIAI, KULCZ2, JACCARD, RAND, ARAND, RR, RT,

/\*SS2, SS1, PHI, DISPER, SEUCLID, MN, OVL, FCA, MI, NMI, VI, AMI, HCV

/asymm= /\*Для асимметричных мер: не выдавать асимметричную версию (NO, тж п/у),

/\*выдать оба асимм значения (BOTH), выдать UV, выдать VU

/normz= /\*Нормирование для NMI, AMI: MAX (тж п/у), MIN, MEAN, GMEAN

/print= /\*Показать технические детали: NO (тж п/у) или YES.

Минимум надо задать VARS, MEASURE.

Макрос вычисляет разные меры согласия между разбиениями, известные как *внешние кластерные критерии* (external clustering criteria), *кластерные валидационные индексы с учителем* (supervised cluster validity indices) и др. названия. Суть в том, чтобы сравнить два разбиения и выдать меру их подобия. Два разбиения подобны, близки, если они состоят из почти одних и тех же групп объектов, т.е. когда верно для многих объектов то, что объекты, образующие группу в одном разбиении, по большей части находятся вместе в группе и в другом разбиении. Макрос вычисляет количественные меры подобия разбиений. Слово «разбиение» это синоним g-классового группирования (классы не пересекаются).

В отличие от макроса [!KO\_CLASAGREE](#_МАКРОС_!CLASAGREE:_МЕРЫ) данный макрос **не требует, чтобы между группами сравниваемых разбиений было известно именное взаимосоответствие**. Т.е. не требуются «метки групп». Это значит, что кодировка групп может быть произвольной и разной в разбиениях, коды не играют роли. Обычно так сравниваются между собой результаты *кластеризаций* или результаты *кластеризации* с *классификацией* или образцовым группированием, сравнение *кластеризации* с *внешним* (фоновым) группированием для валидации кластеризации. Вообще говоря, разбиения могут быть результатами любого по происхождению группирования, причем, как сказано, определенная эквивалентность групп между группированиями не назначена.

Объекты это наблюдения массива, а разбиения это переменные. Их может быть два или больше, макрос сравнивает их по двое. Разбиения должны быть на взаимоисключающие группы ( = альтернативные группы, четкие группы), т.е. каждый объект принадлежит ровно одной группе (группа может состоять и из одного объекта). Данный макрос не вычисляет мер сравнения для неальтернативных (fuzzy) группирований. Этот макрос не использует никак данные-признаки, на основании которых были получены классификации/кластеризации объектов. Только группирование объектов является информацией, входящей в анализ.

Макрос выдает вычисленные меры в окно результатов и в новый массив. Подробнее – см. п/к VARS.

ПРИМЕР 1.

!KO\_cluagree vars= true WITH clu1 clu2 clu3 /measure= RAND ARAND MI AMI.

* Три разных кластерных решения (получены разными методами кластеризации), CLU1, CLU2, CLU3, сравниваются каждое с эталонным разбиением объектов, TRUE, индексами: Rand, Adjusted Rand, Mutual Information, Adjusted Mutual Information.

ПРИМЕР 2.

!KO\_cluagree vars= clu1 to clu10 /measure= FCA.

* Результаты разных кластеризаций (разные методы, разное число кластеров), переменные CLU1 по CLU10, сравниваются друг с другом попарно индексом F Clusteting Accuracy similarity и выдается квадратная матрица.

***Подкоманды***

**VARS**

Укажите входящие числовые категориальные, номинальные переменные (длина имен – до 8 байтов), являющиеся сравниваемыми разбиениями объектов (наблюдений массива данных). Каждое уникальное значение переменной обозначает группу. Переменных минимум две; вы можете использовать слово “to” в списке для смежных в массиве переменных. Переменные должны не иметь пропусков (если есть пропуски, заранее управьтесь с ними: вы можете решить исключить объекты с пропусками или зачислить их в единую новую группу или каждый такой объект в свою группу). Каждая переменная должна не быть константой и должна иметь хотя бы одно повторяющееся значение (т.е. число групп должно быть более одной и меньше числа объектов). Макрос не проверяет корректность входящих переменных, следите за этим сами.

Есть два способа задать VARS: с ключевым словом WITH или же без него. Если без слова WITH, например:

VARS= *CLU1 CLU3* to *CLU10*

Тогда вы имеете право заказать лишь одну меру близости в п/к MEASURE. Макрос вычислит эту меру близости между всеми разбиениями VARS попарно и выдаст как новый массив квадратную матрицу со значениями близости. Каждое разбиение из списка выступает один раз как разбиение *U*, другой раз как разбиение *V*, они сравниваются.

Если использовать слово WITH, то оно должно идти *вторым*, за именем одного разбиения, которое обозначим *U*. Все разбиения, идущие после слова WITH, обозначим как *V*. Одно заданное *U* будет сравнено с каждым из *V*. Например:

VARS= *TRUECLAS* WITH *CLU1 CLU3* to *CLU10*

Разбиение *TRUECLAS* будет сравнено с каждым из идущих после WITH. Это сравнение одного разбиения со всеми остальными. Вы можете затребовать более одной меры близости в п/к MEASURE. Макрос выдаст прямоугольную матрицу результатов в новый массив и в окно результатов.

**ASYMM**

Эта подкоманда касается мер, которые могут быть асимметричными[[1]](#footnote-1) (асимметричными в макросе могут быть следующие меры: DICE, OCHIAI, KULCZ2, OVL, FCA, HCV; подробнее для каждой из них см. п/к MEASURE), т.е. где вычислимы два значения при сравнении разбиений *U* и *V*, *MeasureUV ≠ MeasureVU*. Меры, которые вычисляются лишь симметричными (*MeasureUV = MeasureVU*), игнорируют эту подкоманду.

NO - (тж. по умолчанию/незаданию) выдать симметричную версию меры, которая может быть асимметричной.

UV - выдать значение *MeasureUV*.

VU - выдать значение *MeasureVU*.

BOTH - выдать оба значения, *MeasureUV* (в наддиагональном треугольнике квадратной матрицы, ячейке *UV*) и *MeasureVU* (в поддиагональном треугольнике квадратной матрицы, ячейке *VU*). Этот вариант несовместим с кл. словом WITH в VARS.

**MEASURE**

Выберите меру близости или несколько мер (при наличии слова WITH в списке VARS) из списка. Макрос вычисляет два типа мер.

**A)** Меры на основе **матрицы путаности сочленства** (comembership confusion matrix), или меры **пар объектов** (object pairs). Пусть есть *N* объектов, и два сравниваемых разбиения их на группы, разбиение *U* и разбиение *V*. Рассматриваются все возможные *пары* объектов и наполняется следующая 2x2 частотная таблица (comembership confusion matrix ) подсчета пар:

|  |  |  |  |
| --- | --- | --- | --- |
|  | *V* | |  |
| *U* | *a* | *b* | *a+b* |
| *c* | *d* | *c+d* |
|  | *a+c* | *b+d* | *a+d+b+c = N(N-*1*)/*2 |

по правилу:

* + Если пара находится в одной группе в *U* и находится:
    - * в одной группе в *V* → идет в *a*
      * не в одной группе в *V* → идет в *b*
* Если пара объектов находится не в одной группе в *U* и находится:
  + в одной группе в *V* → идет в *c*
  + не в одной группе в *V* → идет в *d*

Сумма пар *a* является основанием для сходства между *U* и *V*, и сумма пар *d* может быть основанием для сходства в некоторых мерах; *b* и *c* ослабляют сходство.

Следующие меры пар объектов (меры сочленства) в наличии:

DICE - сходство **Дайса**, или **F1**, или **F** мера (**Dice** similarity aka **F1** or **F** measure).

где

называется **точностью** или **определенностью** (**Precision**), а

называется **полнотой** или **воспроизводством** (**Recall**).

*DICE* (F-мера), является среднегармонической воспроизводства *R* и определенности *P*, которые, в ответ, можно назвать двумя «асимметричными версиями» F-меры. При ASYMM=NO выдается, но при ASYMM≠NO макрос выдаст в качестве значение *R*, и в качестве значение *P*.

OCHIAI - сходство **Отиа**, или **Folkes–Mallows** (**Ochiai** similarity aka **Folkes–Mallows**) есть среднегеометрическая воспроизводства *R* и определенности *P*.

При ASYMM=NO выдается, но при ASYMM≠NO макрос выдаст в качестве значение *R*, и в качестве значение *P*.

KULCZ2 - сходство **Кульчинского 2** (**Kulczynski 2** similarity) есть среднеарифметическая воспроизводства *R* и определенности *P*.

При ASYMM=NO выдается, но при ASYMM≠NO макрос выдаст в качестве значение *R*, и в качестве значение *P*.

JACCARD - сходство **Жаккара** (**Jaccard** similarity)

SS2 - сходство **Sokal-Sneath 2** (**Sokal-Sneath 2** similarity)

SS1 - сходство **Sokal-Sneath 1** (**Sokal-Sneath 1** similarity)

RAND - сходство **Рэнда,** или **простого соответствия** (**Rand** aka **Simple matching** similarity)

ARAND - **настроенное** сходство **Рэнда** (**Adjusted Rand** similarity) это *RAND*, нормированное в стохастическом смысле под допущением гипергеометрического распределения. Верхний предел *ARAND* для наблюдаемой пары разбиений равен 1 (если количество групп в разбиениях равное), а к нулю индекс близок в случае, если оба разбиения объектов на группы независимы и случайны (фиксированы лишь объемы групп). *ARAND* может принимать отрицательные значения.

Эта мера тождественна **каппе Коэна** для 2×2 таблицы.

RR - сходство **Рассела–Рао** (**Russel–Rao** similarity)

RT - сходство **Роджерса–Танимото** (**Rogers–Tanimoto** similarity)

PHI - сходство **фи-корреляция** (**Phi correlation** similarity aka **normalized** **Hubert Г** statistic). Это корреляция Пирсона в ситуации двоичных данных. Может колебаться в пределах от -1 до 1.

DISPER - сходство **dispersion** (**Dispersion** similarity). Может колебаться в пределах от -1 до 1.

SEUCLID - различие **квадратное евклидово** расстояние, или **хэммингово** (**squared Euclidean distance** aka **Hamming distance** dissimilarity). Это метрическое расстояние, оно линейно эквивалентно . 2*SEUCLID* известно как **Mirkin distance**.

MN - различие **расстояние Макнемара** (**McNemar distance** dissimilarity). Это квадратный корень из **критерия Макнемара** (McNemar’s test).

**B)** Меры на основе **частотной кросстабуляции** объектов (frequency crosstabulation of objects). Пусть есть *N* объектов, и два сравниваемых разбиения их на группы, разбиение *U* (с группами 1, 2,…, *I*) и разбиение *V* (с группами 1, 2,…, *J*). Имеем кросстабуляцию:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *V* | | | |  |
| *U* |  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |  |

Следующие кросстабуляционные меры в наличии:

OVL - сходство **перекрытие** (**Overlap** similarity aka **Purity**), асимметричная мера при ASYMM≠NO:

и

Ее симметричная версия при ASYMM=NO:

является эквивалентом 1-(**Van Dongen Overlap distance**). 1-*OVL* это метрическое расстояние.

FCA - **F**-сходство **кластеризационной точности** (**F Clusteting Accuracy** similarity), более изощренная мера перекрытия, чем *OVL*, учитывающая вместе понятия «однородности» и «полноты». (Совершенно *однородное* отношение *V* к *U* это когда группа из *V* содержит объекты *только* одной группы из *U*. Совершенно *полное* отношение *V* к *U* это когда группа из *V* содержит *все* объекты некоторой группы из *U*.) *FCA* асимметричная мера, при ASYMM≠NO равная:

и

где это ячейковая F1 мера, равная гармонической средней:

,

Симметричная версия *FCA* при ASYMM=NO:

MI - **взаимоинформационное** сходство (**Mutual information** similarity), энтропийная мера.

есть энтропия разбиения *U*

есть энтропия разбиения *V*

есть совместная энтропия *U* и *V*

Тогда

VI - **взаимоинформационное** расстояние (**Variation of information** distance) это *MI*, переведенное в метрическое расстояние по формуле:

NMI - **нормированное** **взаимоинформационное** сходство (**Normalized Mutual information** similarity) это *MI*, нормированное так, чтобы возможный диапазон меры для *данных* двух разбиений *U* и *V* был от 0 до 1:

, где верхняя граница, , выбирается (подкомандой NORMZ) равной:

- при NORMZ=MAX и по умолчанию,

- при NORMZ=MIN,

- (среднеарифметическая) при NORMZ=MEAN,

- (среднегеометрическая) при NORMZ=GMEAN

AMI - **настроенное** **взаимоинформационное** сходство (**Adjusted Mutual information** similarity) это *MI*, нормированное в стохастическом смысле. Его верхний предел для наблюдаемой пары разбиений равен 1 (как у *NMI*), а к нулю *AMI* близок в случае, если оба разбиения объектов на группы чисто случайны (фиксированы лишь объемы групп). *AMI* может принимать отрицательные значения.

где определено так же, как у *NMI*, а есть ожидаемая величина *MI* при *U* и *V* – двух независимых, случайных распределениях объектов по группам (объемы групп там и там заданы).

(в этой формуле *x* это целое, пробегающие значения «от»[[2]](#footnote-2) и «до»).

HCV - **V**-сходство **однородности и полноты** (**Homogeneity&Completeness V** similarity), энтропийная мера, учитывающая вместе понятия «однородности» и «полноты» (см. *FCA* для сравнения), которые она формулирует так:

, а при

, а при

где условные энтропии (см. меру *MI* выше):

и

Среднегармоническая *h* и *c*, симметричная мера, есть *HCV*:

Если затребована асимметричная мера (ASYMM≠NO), то *c* и *h* выдаются вместо :

и

Источники:

* Wagner, S., Wagner, D. Comparing clusterings - an overview // *A manuscript: Univ., Fak. für Informatik*, Janyary 12, 2007.
* Vinh, N.X., Epps, J., Bailey, J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance // *Journal of Machine Learning Research*, 2010, 11, 2837-2854.
* Vinh, N.X., Epps, J., Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? // *Proceedings of the 26 th International Conference on Machine Learning*. Montreal, Canada, 2009.
* Rosenberg, A., Hirschberg, J. V-measure: a conditional entropy-based external cluster evaluation measure // *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Prague, June 2007, 410–420.
* Romano, S. et al. Adjusting for chance clustering comparison measures // *Journal of Machine Learning Research*, 2016, 17, 1–32.
* Meila, M. Comparing clusterings – an information based distance // *Journal of Multivariate Analysis*, 2007, 98, 873–895.
* Desgraupes, B. Clustering indices // *Package clusterCrit for R*, April 2013.
* Rendon, E. et al. Internal versus external cluster validation indexes // *International Journal of Computers and Communications*, 2011, 1 (1), 27–34.
* Halkidi, M., Batistakis, Y., Vazirgiannis, M. On clustering validation techniques // *Journal of Intelligent Information Systems*, 2001, 17 (2/3), 107–145.

**NORMZ**

Эта подкоманда нужна для вычисления мер NMI и AMI (см. п/к MEASURE), она задает метод нормирования. Выберите одно из: MAX (тж. по умолчанию), MIN, MEAN, GMEAN.

**PRINT**

Задайте PRINT=YES, если хотите видеть некоторые технические детали, а именно значения A, B, C, D из таблиц путанности сочленства (из них вычисляются меры типа “A”) и частотные кросстабуляции (из них вычисляются меры типа “B”). В частотных кросстабуляциях значения переменных (коды групп) макрос не отображает, а порядок их такой, в каком значения были встречены в данных, а не сортированный. По умолчанию PRINT=NO.

***Особые режимы***

Макрос игнорирует взвешивание (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива данных. Макрос слушается команд, выбирающих наблюдения (FILTER, USE, SELECT IF, N OF CASES) и преобразований, в том числе под командой TEMPORARY.

# МАКРОС !KO\_CLASAGREE: МЕРЫ СРАВНЕНИЯ РАЗБИЕНИЙ-«КЛАССИФИКАЦИЙ»

Version 1, Mar 2019. Tested on SPSS Statistics 17, 20, 22.

!KO\_clasagree vars= v1 WITH v2 to v6 /\*Переменные, являющие собой группирования наблюдений;

/\*слово WITH возможно после первого имени

/measure= ACC F1 OSR /\*Укажите меру(ы): ACC, REC, PRE, SPE, NPV, YOUD, MARK,

/\*F1, FBETA, KULCZ2,

/\*LNDOR, DP, CORR, BCR, GM, AGM, OPRE, JACCARD, OSR, KAPPA, SPI, BPK, RKCORR

/binar= /\*Как поступитьс двоичными мерами: CLASS, MEAN (тж п/у), WMEAN, IWMEAN

/beta= 0.5 /\*Beta для меры FBETA

/print= YES /\*Показать технические детали: NO (тж п/у) или YES.

Минимум надо задать VARS, MEASURE.

Макрос вычисляет разные меры согласия между разбиениями с «помеченными» (идентифицированными) группами. Эти меры известны как *индексы согласия классификаций*, меры *исполнительского качества классификаторов* и др. названия. В сравниваемых разбиениях *U* и *V* **есть заведомо взаимосоответствующие («те же самые»), по имени, группы**. Это отражается в кодировке данных, которая должна быть единообразной. Группа с кодом, допустим, 1 в *U* – это «та же» группа, что группа с кодом 1 в *V*; группа с кодом 2 в *U* – это «та же» группа, что группа с кодом 2 в *V*, и т.д. Следовательно, *U* и *V* хорошо согласуются (подобны), если преобладают объекты, попадающие в них в группы с одним и тем же кодом: 1 в *U* и 1 в *V*, 2 в *U* и 2 в *V*. Слово «разбиение» это синоним g-классового группирования (классы не пересекаются).

Если же взаимосоответствующие группы в разбиениях не идентифицированы, т.е. неизвестно, какой группе в *U* отвечает группа под кодом 1 (например) в *V*, тогда используйте макрос [!KO\_CLUAGREE](#_МАКРОС_!PROXQNT:_РАЗНЫЕ). Или установите сначала взаимосоответствующие группы каким-нибудь путем, к примеру, эмпирически, макросом [!KO\_GRMATCH](#_МАКРОС_!GRMATCH:_ИДЕНТИФИКАЦИЯ_1).

По происхождению разбиения могут быть любыми для макроса !KO\_CLASAGREE: результаты классификаций, кластеризаций, естественные или образцовые группирования. Совпадение кодов между разбиениями может быть не полным и число групп в разбиениях может быть не одинаково. Скажем, *U* может состоять из групп 1, 2 и 3, а *V* – из групп 1, 2, 4 и 5. В данном случае взаимосоответствующие группы – 1 и 2, они есть в обоих разбиениях; остальные группы – невзаимосоответстующие и не могут быть основанием для согласия между *U* и *V*, наоборот – они голосуют за несогласие. Обыкновенно же при сравнении классификаций используют, однако, данные, где *U* и *V* сложены лишь из взаимосоответствующих групп.

Объекты это наблюдения массива, а разбиения это переменные. Их может быть два или больше, макрос сравнивает их по двое. Разбиения должны быть на взаимоисключающие группы ( = альтернативные группы, четкие группы), т.е. каждый объект принадлежит ровно одной группе (группа может состоять и из одного объекта). Данный макрос не вычисляет мер сравнения для неальтернативных (fuzzy) группирований (например, multi-topic классификаций). Этот макрос не использует никак данные-признаки, на основании которых были получены классификации/кластеризации объектов. Только группирование объектов является информацией, входящей в анализ.

Макрос выдает вычисленные меры в окно результатов и в новый массив. Подробнее – см. п/к VARS.

ПРИМЕР 1.

!KO\_clasagree vars= true WITH clas1 clas2 clas3 /measure= ACC F1 OSR.

* Три разных разбиения, полученные разными классификаторами, CLAS1, CLAS2, CLAS3, сравниваются каждое с эталонной классификацией, TRUE, индексами: Accuracy, F1-measure, Relative Agreement (OSR).

ПРИМЕР 2.

!KO\_clasagree vars= clas1 to clas10 /measure= KAPPA.

* Результаты разных классификаций (разные эксперты, классифицировавшие объекты по группам), переменные CLAS1 по CLAS10, сравниваются друг с другом попарно каппой Коэна и выдается квадратная матрица.

***Подкоманды***

**VARS**

Укажите входящие числовые категориальные, номинальные переменные (длина имен – до 8 байтов), являющиеся сравниваемыми разбиениями объектов (наблюдений массива данных). Каждое уникальное значение переменной обозначает группу. Переменных минимум две; вы можете использовать слово “to” в списке для смежных в массиве переменных. Переменные должны не иметь пропусков (если есть пропуски, заранее управьтесь с ними: вы можете решить исключить объекты с пропусками или зачислить их в единую новую группу или каждый такой объект в свою группу). Каждая переменная должна не быть константой и должна иметь хотя бы одно повторяющееся значение (т.е. число групп должно быть более одной и меньше числа объектов). Макрос не проверяет корректность входящих переменных, следите за этим сами.

Есть два способа задать VARS: с ключевым словом WITH или же без него. Если без слова WITH, например:

VARS= *CLAS1 CLAS3* to *CLAS10*

Тогда вы имеете право заказать лишь одну меру близости в п/к MEASURE. Макрос вычислит эту меру близости между всеми разбиениями VARS попарно и выдаст как новый массив квадратную матрицу со значениями близости. Если мера является симметричной, на выходе будет полная матрица. Если мера является асимметричной, на выходе будет матрица с заполненным наддиагональным треугольником. Ячейка (i,j) в матрице содержит результат сравнения i-й переменной в качестве разбиения *U* с j-й переменной в качестве разбиения *V*. См. ПРИМЕР 4, если вам надо заполнить поддиагональный треугольник тоже. На диагонали стоит 1, если верхний предел близости единица, и стоит произвольное число 999, если верхний предел близости +∞.

Если использовать слово WITH, то оно должно идти *вторым*, за именем одного разбиения, которое обозначим *U*. Все разбиения, идущие после слова WITH, обозначим как *V*. Одно заданное *U* будет сравнено с каждым из *V*. Если вы сравниваете эталонную классификацию с одной или несколькими предсказанными классификациями, то эталонная должна быть *U*. Например:

VARS= *TRUECLAS* WITH *CLAS1 CLAS3* to *CLAS10*

Разбиение *TRUECLAS* будет сравнено с каждым из идущих после WITH. Это сравнение одного разбиения со всеми остальными. Вы можете затребовать более одной меры близости в п/к MEASURE. Макрос выдаст прямоугольную матрицу результатов в новый массив и в окно результатов.

**MEASURE**

Выберите меру близости или несколько мер (при наличии слова WITH в списке VARS) из списка. Макрос вычисляет два типа мер: меры **для двоичной** классификации (значения которых макрос может усреднить в значения для мультиклассовой классификации) и меры **собственно для мультиклассовой** классификации.

И те и другие меры исходят из частотной кросстабуляции сравниваемых разбиений (классификаций) *U* (*I* групп) и *V* (*J* групп):

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | *V* | | | |  |
| *U* |  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |  |

Рис. 1.0

Имена (метки) классов известны, и они частью или полностью одинаковы в *U* и *V*. (Если *U* и *V* не имеют ни одного общего класса, макрос не вычислит мер для этой пары разбиений и уведомит о том.) Вот примеры кросстабуляций:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | *V* | | | |
|  |  | *class 1* | *class4* | *class 2* | *class 3* |
| *U* | *class 1* | 12 | 0 | 6 | 3 |
| *class 3* | 7 | 10 | 5 | 8 |
| *class 2* | 2 | 1 | 21 | 0 |

Рис. 1.1

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | *V* | |
|  |  | *class 1* | *class 2* |
| *U* | *class 3* | 1 | 2 |
| *class 2* | 9 | 28 |
| *class 4* | 3 | 10 |
| *class 1* | 35 | 0 |

Рис. 1.2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | *V* | | |
|  |  | *class 1* | *class 2* | *class 3* |
| *U* | *class 1* | 42 | 11 | 7 |
| *class 2* | 0 | 14 | 1 |
| *class 3* | 18 | 5 | 26 |

Рис. 1.3

Ячейки на пересечении одноименных классов (т.е. групп, кодированных в переменных *U* и *V* одним и тем же кодом) в этих примерах помечены. Заметьте, что не обязательно в общем случае, чтобы *U* и *V* состояли полностью из одних и тех же классов, как это на рис. 1.3, где оба разбиения исчерпываются классами 1, 2, 3. На рис. 1.1 *U* состоит из трех классов, а *V* состоит из четырех, три класса (условно обозначенные 1, 2, 3) – общие у них. Аналогично, на рис. 1.2 *U* состоит из четырех классов, а *V* – из двух, эти два класса общие для *U* и *V*. Сортировка рядов и столбцов кроссчастотки по кодам классов не играет роли в анализе. Так, на рис. 1.1 и 1.2 порядок классов не сортирован по возрастанию. Краевые рядные и столбцовые частоты всегда ненулевые, поскольку макрос производит таблицы на основании только наличных данных. Минимальный размер таблицы 2x2.

Обозначим набор классов, общих для *U* и *V*, буквой *M*, а число этих общих классов буквой *K*, 1≤*K*≤min(*I,J*). *N* же это суммарная частота всей таблицы (= число наблюдений массива данных), а не только *K* помеченных в ней ячей.

Меры согласия (подобия) между *U* и *V* бывают симметричными и асимметричными. *Симметричная* мера не изменится, если *U* (ряды таблицы) и *V* (столбцы таблицы) поменять местами, т.е. транспонировать таблицу частот. *Асимметричная* мера изменит свое значение при транспонировании таблицы[[3]](#footnote-3). **Для асимметричных мер имеет значение**, какое из двух разбиений считать *эталонным* (образцовым, подлинным), а какое считать *предсказанным* (опытным). Макрос !KO\_CLASAGREE **трактует разбиение *U* (задающее ряды таблицы) эталонной классификацией, а разбиение *V* (задающее столбцы таблицы) предсказанной классификацией**.

**A**) Меры для **двоичной** (**унитарной, класс-специфичной**) классификации (**binary** aka **one-hot** aka **class-specific** classification). Интерес сфокусирован на классифицировании в один особый класс.

Такая мера подобия (согласия) вычисляется для каждого класса *k* из *M*. Макрос строит для него следующую **2x2 матрицу путанности** (2x2 confusion matrix), где разбиение *U* (ряды) принимается за эталонную классификацию, а разбиение *V* (столбцы) принимается за предсказанную классификацию:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | *V* | |  |
|  |  | *class k* | *не class k* |  |
| *U* | *class k* | True Positives  *a* | False Negatives  *b* | *a+b* |
| *не class k* | False Positives  *c* | True Negatives  *d* | *c+d* |
|  |  | *a+c* | *b+d* | *a+d+b+c = N* |

Эта матрица есть дихотомизованная кросстабуляция *I*x*J*. На основании частот *a*, *b*, *c*, *d* вычисляется та или иная мера. Т.к. матриц путанности *K* штук, получается *K* значений меры. Ниже даны формулы разных мер. Подкоманда BINAR=CLASS выдает все *K* значений, одно для каждого класса из *M*. При ином задании подкоманды BINAR (см. там) *K* значений усредняются в одно значение, характеризующее согласие между *U* и *V* в целом.

Некоторые двоичные меры симметричны, другие асимметричны – при BINAR=CLASS или MEAN (симметричные меры не меняют значение, если *U* и *V* поменять местами). При BINAR=WMEAN или IWMEAN все двоичные меры становятся асимметричны, поскольку взвешивание всегда делается объемами классов того разбиения, которое образует *ряды* кросстабуляции и матрицы путанности. Следующие двоичные меры в наличии (все они – сходства):

ACC - **точность**,или **коэффициент Рэнда** (**Accuracy** aka **Rand** aka **Simple Matching** coefficient), симметрична, варьирует [0,1].

1-*ACC* называется **Missclassification** (или **Error**) **Rate** и линейно эквивалентно квадратному евклидовому расстоянию.

REC - **воспроизводство**, или **чувствительность** (**Recall** aka **Sensitivity** aka **True Positive Rate** aka **Hit Rate** aka **Positive Accuracy**), асимметрична (симметричное значение =*PRE*), варьирует [0,1].

SPE - **специфичность** (**Specificity** aka **True Negative Rate** aka **Negative Accuracy**), асимметрична (симметричное значение =*NPV*), варьирует [0,1].

YOUD - **Youden's index**, асимметрична, варьирует [-1,1].

PRE - **определенность** (**Precision** aka **Positive Predictive Value**), асимметрична (симметричное значение =*REC*), варьирует [0,1].

1-*PRE* называется **False Discovery Rate.**

NPV - **Negative Predictive Value**, асимметрична (симметричное значение =*SPE*), варьирует [0,1].

MARK - **Markedness index**, асимметрична, варьирует [-1,1].

F1 - **F1** или **F мера**, или **коэффициент Дайса** (**F1** aka **F Measure** aka **Dice Matching** coefficient), симметрична, варьирует [0,1], является среднегармонической *REC* и *PRE*.

При нулевых *REC* и *PRE* мера не вычисляется.

FBETA - **F-beta мера**, или **взвешенная F мера** (**F-beta** aka **generalized** or **weighted F Measure**), в общем асимметрична, варьирует [0,1], является взвешенной среднегармонической *REC* и *PRE*.

,

где параметр *beta* (0,+∞) задается подкомандой BETA. Если *beta*<1, *PRE* получает больший вес, чем *REC*; если *beta*>1, то наоборот, *REC* получает больший вес. При *beta*=1 *FBETA* превращается в *F1*. При нулевых *REC* и *PRE* мера не вычисляется.

KULCZ2 - **коэффициент Кульчинского 2** (**Kulczynski 2** coefficient), симметрична, варьирует [0,1], является среднеарифметической *REC* и *PRE*.

2*KULCZ2*-1 называется **Classification Success Index**.

LNDOR - **логарифм Diagnostic Odds Ratio**, симметрична, варьирует [-∞,+∞].

Если *a*, *b*, *c* или *d* нуль, мера не вычисляется.

DP - **Discriminant Power**, симметрична, варьирует [-∞,+∞].

Если *REC* или *SPE* равно 1 или 0, мера не вычисляется.

CORR - **бинарная корреляция** **Мэттьюса** (**Matthews correlation** aka **Phi correlation)**, симметрична, варьирует [-1,1]. Это ни что иное как коэффициент корреляции Пирсона в случае двоичных данных.

,

где и .

BCR - **Balanced Classification Rate**, асимметрична, варьирует [0,1], является среднеарифметической *REC* и *SPE*; это AUC (площадь под кривой) в ROC-пространстве, для единственной точки там.

GM - **GM Measure**, асимметрична, варьирует [0,1], является среднегеометрической *REC* и *SPE*.

AGM - **Adjusted GM Measure**, асимметрична, варьирует [0,1], является модификацией *GM*, созданной, чтобы лучше справляться с несбалансированными классами.

, а при *REC*=0, *AGM*=0

where

Если позитивный двоичный класс диспропорционально мал, эта мера может быть предпочтительнее *GM*, поскольку она чувствительнее к изменениям *SPE*, чем к изменениям *REC*.

OPRE - **Optimized Precision**, асимметрична, варьирует [-∞,1]. Мера выше тогда, когда *a* и *d* оба высоки.

Если *REC*+*SPE*=0, мера не вычисляется.

JACCARD - коэффициент **Жаккара**, или **Танимото** (**Jaccard** aka **Tanimoto** Matching coefficient), симметрична, варьирует [0,1].

Меры ACC, PRE, NPV, CORR, F1, FBETA, OPRE, JACCARD чувствительны к изменению формы распределения объектов по классам эталонной (*U*) классификации, поэтому говорят, что эти меры не одинаковы в случаях сбалансированных и несбалансированных классов.

**B**) Меры для **собственно мультиклассовой** (**номинальной**) классификации (**multiclass proper** aka **nominal** classification). Это симметричные меры ассоциации двух номинальных переменных, известные также как меры согласия экспертов.

Такая мера, ее единое значение, вычисляется непосредственно из всей частотной кросстабуляции (см. рис. 1.0–1.3). Основанием близости являются частоты в ячейках на пересечении одноименных классов, т.е. это *K* ячеек, составляющих набор *M*. Следующие мультиклассовые меры в наличии, все они сходства:

OSR - **Общая доля успеха** или **относительное согласие** или **мультиклассовая точность** (**Overall Success Rate** aka **Relative Agreement** aka **Multiclass Accuracy**) это просто доля одинаково классифицированных объектов, т.е. доля объектов, которые в *U* и *V* находятся в группах одного и того же кода.

KAPPA - **Каппа Коэна** (**Cohen’s Kappa**) это *OSR*, нормированное уровнем случайного согласия. Варьирует [-∞, 1]. Значение 0 означает согласие на уровне случайного.

, где наблюдаемое согласие , а ожидаемое случайное согласие

SPI - **Пи Скотта** (**Scott’s Pi**) это другой способ нормировать *OSR* уровнем случайного согласия. Варьирует [-∞, 1]. Значение 0 означает согласие на уровне случайного. Формула *SPI* та же, что у *KAPPA*, но

BPK - **Каппа** **Brennan–Prediger** – еще способ нормировать *OSR* уровнем случайного согласия, без допущения фиксированных объемов классов (т.е. краевых частот). Варьирует [-∞, 1]. Значение 0 означает согласие на уровне случайного. Формула *BPK* та же, что у *KAPPA*, но

Если *I=J=K*, то и *BPK* линейно эквивалентна *OSR*.

RKCORR - **мультиклассовая корреляция Мэттьюса** (**Rk корреляция**). Это корреляция Пирсона, посчитанная биективно между двумя наборами фиктивных (dummy) переменных, варьирует [-1, 1]. Если превратить две категориальные переменные, состоящие из одинаковых *K* категорий, в два набора двоичных фиктивных переменных по *K* столбцов, где столбцы – они соответствуют категориям – между наборами соупорядочены, то после центрации всех столбцов надо векторизовать каждый *N*x*K* набор в столбец длиной *NK*. Тогда r Пирсона, посчитанный по «формуле косинуса» между двумя полученными столбцами X и Y, есть *RKCORR*:

, где это сумма кросс-произведений X и Y, а и это суммы квадратов в двух столбцах.

Эквивалентная формула вычисления *RKCORR* из квадратной *K*x*K* кросстабуляции с одинаковыми, соупорядоченными категориями (см. рис. 1.3 выше):

*RKCORR* не вычисляется, если разбиения *U* и *V* состоят из не вполне одинаковых наборов групповых кодов.

Источники:

* Tharwat, A. Classification assessment methods // *Applied Computing and Informatics* (2018), https://doi.org/10.1016/j.aci.2018.08.003.
* Labatut, V., Cherifi, H. Evaluation of performance measures for classifiers comparison // *Ubiquitous Computing and Communication Journal,* 2011, 6, 21-34.
* Hossin, M., Sulaiman, M. A review on evaluation metrics for data classification evaluations // *International Journal of Data Mining & Knowledge Management Process (IJDKP),* 2015, 5(2), 1-11.
* Sokolova, M., Lapalme, G. A systematic analysis of performance measures for classification tasks // *Information Processing and Management*, 2009, 45, 427–437.
* Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme // *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 1975, 405(2), 442–451.
* Gorodkin, J. Comparing two K-category assignments by a K-category correlation coefficient // *Computational Biology and Chemistry*, 2004, 28, 367-374.
* Jurman, G., Riccadonna, S,, Furlanello, C. A comparison of MCC and CEN error measures in multi-class prediction // *PLoS ONE*, 2012, 7(8): e41882, doi:10.1371/journal.pone.0041882.
* Brennan, R., Prediger, D. Coefficient Kappa: some uses, misuses and alternatives // *Educational and Psychological Measurement*, 1981, 41, 687-699.
* Maxwell, A.E. Coefficients of agreement between observers and their interpretation // *Brit. J. Psychiatry*, 1977, 530, 79-83.
* Batuwita, R., Palade, V. Adjusted geometric-mean: a novel performance measure for imbalanced bioinformatics datasets learning // *Journal of Bioinformatics and Computational Biology*, 2012, 10(4), 1250003-1­–1250003-23.

**BINAR**

Эта подкоманда ­– для мер двоичного сравнения (см. п/к MEASURE, меры “**A**”), она игнорируется мультиклассовыми мерами (меры “**B**”). При сравнении двух разбиений двоичные меры вычисляются сначала для каждого из *K* общих классов. BINAR=CLASS выводит в результаты эти поклассовые значения. BINAR=CLASS несовместимо с заданием каких-либо мультиклассовых мер. BINAR=CLASS позволено только при использовании WITH в VARS.

Другие опции BINAR подразумевают усреднение *в процессе* или *после* вычисления меры (см. ниже), и таким образом макрос выведет одно значение, характеризующее степень согласия между двумя разбиениями. Способы усреднения:

MEAN - (тж по умолчанию/незаданию) простая средняя арифметическая *K* значений.

WMEAN - взвешенная средняя средняя арифметическая *K* значений. Весами выступают *рядные* краевые частоты, т.е. объемы классов разбиения *U*. Таким образом, классы большего объема сказываются на усредненном результате сильнее. Используйте, если вам важнее качество предсказания для больших классов.

IWMEAN - взвешенная средняя арифметическая *K* значений. Подобно предыдущему, но весами выступают обращенныерядные краевые частоты, так что классы меньшего объема сказываются на усредненном результате сильнее. Используйте, если вам важнее качество предсказания для малых классов.

При BINAR= WMEAN и IWMEAN все меры двоичного сравнения становятся асимметричными, поскольку весами выступают всегда объемы в *эталонной* классификации (то, что задает *ряды* кросстабуляции).

*Когда происходит усреднение*. У следующих мер *K* значений усредняются *после* того, как мера вычислена: ACC, REC, PRE, SPE, NPV, LNDOR, DP, JACCARD. Усредняются *K* значений меры.

У следующих мер усреднение делается *перед* вычислением меры по своей формуле: YOUD, MARK, F1, FBETA, KULCZ2, CORR, BCR, GM, AGM, OPRE. У этих мер усредняются *K* значений каждого *члена* формулы, и затем уже одно значение меры вычисляется по ее формуле. Члены, на уровне которых делается усреднение, вы можете видеть в формулах (см. п/к MEASURE). Для большинства этих мер членами формул выступают: *REC*, *PRE*, *SPE*, *NPV*. Например, для получения усредненной меры F1 сначала усредняется *K* значений *PRE*, *K* значений *REC*. Полученные средние *PREM* и *RECM* подставляются в формулу F1. У меры CORR членами формулы, подвергаемыми усреднению, являются *a/N*, *S* и *P*; у меры AGM – *REC*, *SPE* и *Q*.

Заметим, что усредненная двоичная мера CORR это не то же самое, что мультиклассовая мера RKCORR, хотя и та и эта основаны на пирсоновской корреляции.

Мера DP имеет членами *REC* и *SPE*, тем не менее она усредняется после формулы, а не перед формулой; так сделано в интересах поддержания ее симметричности.

ПРИМЕР 3.

!KO\_clasagree vars= class WITH clas1 clas2 /measure= REC PRE SPE NPV /binar=CLASS /print=YES.

* Опытные классификации CLAS1 и CLAS2 оцениваются относительно идеальной классификации CLASS. Заказаны, для каждого существующего класса, основные асимметричные статистики этой оценки. Также затребовано показать частоты, на которых они вычислены (PRINT=YES).

**BETA**

Эта подкоманда используется только мерой FBETA(см. формулу). Укажите в качестве весового параметра *beta* число выше нуля. Если *beta*<1, *PRE* получает больший вес, чем *REC*; если *beta*>1, то наоборот, *REC* получает больший вес. При *beta*=1 мера FBETA превращается в меру F1. Обычно задают значения 2 или 0.5.

**PRINT**

Задайте PRINT=YES, если хотите видеть некоторые технические детали, а именно частотные кросстабуляции. В частотных кросстабуляциях порядок классов не сортированный, такой, в каком значения были встречены в данных. Если заказаны меры двоичного сравнения, макрос показывает также все поклассовые частоты TP(*a*), FN(*b*), FP(*c*), TN(*d*). По умолчанию PRINT=NO.

ПРИМЕР 4. Получение обоих матричных треугольников для асимметричной меры. Анализируется массив Data.

!KO\_clasagree vars= p1 p2 p3 p4 p5 /measure= BCR.

dataset name mx1.

* BCR асимметричная мера, поэтому на выходе в матрице заполнен только наддиагональный треугольник. В ячейке (1,2), к примеру, содержится значение BCRP1P2, где P1 было разбиением U (задававшим ряды кросстабуляции), а P2 было разбиением V (задававшим столбцы кросстабуляции). Мы хотим получить симметричные значения, например BCRP2P1, которые заполнили бы собой поддиагональный треугольник матрицы. Прием такой:

compute casenum= $casenum.

dataset activate Data.

!KO\_clasagree vars= p5 p4 p3 p2 p1 /measure= BCR.

compute casenum= $casenum.

sort cases by casenum (D).

compute casenum= $casenum.

execute.

update file= mx1 /file= \* /by casenum.

execute.

delete var casenum.

dataset close mx1.

* Сначала создали в выведенной матрице переменную-индикатор наблюдений CASENUM. Затем вернулись к Data и пустили макрос еще раз, но со списком VARS, написанным задом наперед. Макрос выдал искомые значения, однако они находятся опять в верхнем треугольнике. Поэтому создали и тут индикатор CASENUM и сортировали им наблюдения в обратном порядке. Сделав это, опять пересчитали CASENUM (таким образом последнее наблюдение, ставшее после сортировки первым, имеет CASENUM=1). После этого командой UPDATE объединили данные этой матрицы с первой матрицей (MX1). В результате имеем правильно заполненные оба треугольника матрицы.

***Особые режимы***

Макрос игнорирует взвешивание (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива данных. Макрос слушается команд, выбирающих наблюдения (FILTER, USE, SELECT IF, N OF CASES) и преобразований, в том числе под командой TEMPORARY.

# МАКРОС !KO\_GRMATCH: ИДЕНТИФИКАЦИЯ (СПАРИВАНИЕ) ВЗАИМОСООТВЕТСТВУЮЩИХ ГРУПП

Version 1, Mar 2019. Tested on SPSS Statistics 17, 20, 22.

!KO\_grmatch vars= v1 WITH v2 to v6 /\*Переменные, являющие собой группирования наблюдений;

/\*слово WITH возможно после первого имени

/matchon= /\*Спаривать на базе частот (COUNT, тж п/у), остатков (RESID)

/\*или F-мер (FCELL)

/method= GREEDY /\*Алгоритм спаривания: Венгерский (HUNGAR), жадный (GREEDY)

/\*или жадный с запретом (PGREEDY)

/recode= 'd:\exercise\RecSyntax.sps' /\*Перекодировать ли переменные: NO (тж п/у),

/\*YES или путь/имя sps-файла

/other= /\*Неспаренные группы: оставить отдельными (INDIVID, тж п/у) или объединить

/\*(задайте код группы)

/simmx= /\*Создать матрицу спаренностей групп и показать на дендрограмме: YES

/\*или NO (тж п/у)

/print= /\*Отчитываться о спаривании в окно результатов: YES (тж п/у) или NO.

Минимум надо задать VARS, METHOD.

Допустим, у вас есть массив наблюдений и два или более разбиения этого массива на группы наблюдений (объектов). Разбиения могут быть результатами кластеризации, классификации, иного по происхождению группирования (в том числе естественное или подлинное/идеальное группирование). Вы ожидаете, что разбиения одни и те же или, во всяком случае, довольно похожи – в том смысле, что наблюдения, находящиеся в одной группе в одном разбиении, преимущественно находятся в одной группе и в другом разбиении. Но проблема данных состоит в том, что коды, “метки групп”, неодинаковы в разбиениях. Например, разбиения *U* и *V* одинаковы или похожи и состоят из групп A, B и C, но эти группы *кодированы* в *U* значениями 1, 2, 3, а в *V* соответственно значениями 2, 6, 4. Вам *неизвестно*, *какая* группа в *V* соответствует *какой* группе в *U*, соответствует **в том смысле, что состоит преимущественно из тех же наблюдений массива**. Идентификация этого, эмпирического соответствия – взаимного соответствия между группами разных разбиений по их наполненности – и есть задача данного макроса. Слово «разбиение» это синоним g-классового группирования (классы не пересекаются).

Разбиения это категориальные, номинальные переменные. Каждое наблюдение принадлежит ровно одной группе (классу, кластеру и т.п.) в каждом разбиении, т.е. разбиение это одна переменная массива и группы в нем взаимоисключающие. Группа может состоять и из одного объекта. Задача макроса – сопоставить попарно разбиения на предмет того, в какие группы попадают одни и те же наблюдения в них. Если преимущественно те же самые наблюдений попадают в группу A разбиения *U*, что попадают в группу X разбиения *V*, то A и X – это примерно «одна и та же» группа, т.е., другими словами, A и X взаимосоответствующи. Пускай *U* содержит 3 группы, а *V* содержит 5 групп. Макрос спарит три группы *U* с тремя группами *V*; это *взаимно-однозначное* спаривание (one-to-one pairing): одна группа – одна группа. Две «лишние» группы *V*, которые макрос сочтет недостаточно похожими на группы *U*, останутся не спарены.

Переменные (разбиения) сопоставляются по две. Спаривание групп (т.е. идентификация взаимосоответствующих) между ними делается на основе их кросстабуляции (частоты, остатки или F-меры) алгоритмом спаривания, старающимся максимизировать сумму в выбранных парах.

Макрос может проделать или выписать как синтаксис-файл перекодировку переменных, чтобы у них одинаковые/сходные группы были кодированы одинаково (имели одну и ту же “метку”). Вы можете также увидеть группы, признанные макросом взаимосоответствующими, на дендрограмме.

***Подкоманды***

**VARS**

Задайте список категориальных числовых переменных (длина имен – до 8 байтов), вы можете использовать “to” в списке. Пропуски не разрешены в этих данных. Каждая переменная это разбиение наблюдений массива на группы. Разные переменные – это разные такие разбиения. Макрос будет сопоставлять переменные по две и устанавливать, какие группы одного разбиения соответствуют каким группам второго разбиения. Каждое уникальное значение в переменной обозначает одну группу: переменные являются номинальными. Неважно, одинаковы или неодинаковы значения, коды групп, у разных переменных списка.

*Допустимая кодировка*. Значения могут иметь любой знак, быть целыми или дробными. Ширина значения (включая отрицательный знак и десятичный разграничитель, если есть) – до 8 цифр. Если код дробный, он должен иметь *не более двух* значимых десятичных цифр. Макрос не проверяет входящие значения, следите сами, чтобы они не нарушали описанные требования.

*Слово WITH*. После *первого* имени в VARS можно вставить однократно ключевое слово WITH. Тогда каждая из остальных переменных VARS будет сопоставлена с одной переменной – той, что упомянута перед WITH. Если слова WITH в списке нет, тогда все переменные будут сопоставлены друг с другом попарно.

ПРИМЕР 1.

!KO\_grmatch vars= true WITH clu1 clu2 clu3 /method= GREEDY /recode= 'd:\exercise\RecSyntax.sps'.

* Три разных кластерных решения (получены разными методами кластеризации), CLU1, CLU2, CLU3, сопоставляются каждое с эталонной классификацией TRUE. Жадный метод спаривания применяется для установления взаимосоответствующих групп. Затребовано выписать синтаксис для перекодировки переменных CLU1, CLU2, CLU3, чтобы сделать коды групп в них единообразными с таковыми переменной TRUE.

ПРИМЕР 2.

!KO\_grmatch vars= p1 to p8 /method= HUNGAR /matchon= FCELL /simmx= YES /print= NO.

* Все альтернативные группирования с P1 по P8 сопоставляются друг с другом попарно в поиске взаимосоответствующих групп между ними. Поиск осуществляется Венгерским алгоритмом спаривания на базе F-мер. Результаты – «одни и те же» группы разных группирований затребовано показать графически на дендрограмме. Выдача подробных результатов отменена.

**MATCHON**

Выберите, какую таблицу показывать в отчете как «Кросстабуляцию» и на базе чисел которой, значений в ячейках *ij*, делать спаривание между ее рядами (группы одной переменной) и ее столбцами (группы другой переменной):

COUNT - (тж по умолчанию/незаданию) таблица частот: . Это значит, что за признак взаимосоответствия вы принимаете высокую, по сравнению со среднетабличной, наблюдаемую частоту в ячейке.

RESID - таблица частотных остатков: . Это значит, что за признак взаимосоответствия вы принимаете частоту в ячейке, которая высока по сравнению с ожидаемой частотой для данной ячейки. Ожидаемая частота ячейки это частота, полностью определяемая парой ее краевых частот и , т.е. объемами двух рассматриваемых групп. Если наблюдаемое пересечение двух групп велико, но велико оно только лишь оттого, что сами эти группы объемные, то это не становится слишком весомым доводом за их взаимосоответствие.

FCELL - таблица ячейковых F-мер: . Это еще один способ оперировать не самими частотами, а их отношением с краевыми частотами. F-мера в ячейке является среднегармонической ячейковых “полноты” , и “точности” .

**METHOD**

Выберите алгоритм спаривания.

HUNGAR - Венгерский (Kuhn–Munkres) алгоритм. Алгоритм итерационный, он максимизирует сумму значений в выбранных ячейках (т.е. в спаренных рядах и столбцах) до глобального оптимума.

GREEDY - простой жадный алгоритм. Шаговый, он быстрее, но не гарантирует достижения глобального оптимума, т.е. наибольшей возможной суммы значений в выбранных ячейках, хотя и нацелен максимизировать ее. Алгоритм выбирает на каждом шаге наибольший сейчас элемент в таблице, считает его сделанным спариванием и удаляет из таблицы этот ряд и этот столбец. Этот подход с первых шагов выбирает самые большие значения, тогда как Венгерский более «осторожный» и может иногда предпочесть не самый большой элемент, если в итоге это даст бо̀льшую сумму.

PGREEDY - жадный алгоритм с запретом. Это то же, что GREEDY, но накладывается дополнительное условие: наибольший сейчас элемент в таблице считается сделанным спариванием, если он есть максимальный в своем ряду и в своем столбце в исходной таблице. В противном случае спаривание на шаге не засчитывается, хотя ряд и столбец из таблицы удаляются. Этот вариант, таким образом, разрешает спаривание только за счет самых больших значений в рядах и в столбцах кросстабуляции. Он является ни чем иным как GREEDY, в котором «малоубедительные» спаривания просто отменены.

Если в сравниваемых двух разбиениях число групп *I* и *J*, то HUNGAR и GREEDY образуют всегда *min(I,J)* пар, но PGREEDY может дать меньше пар.

**RECODE**

Эта подкоманда позволяет перекодировать переменные VARS, следующие после слова WITH, в соответствие переменной, стоящей перед WITH. Перекодировка диктуется соответствием групп, найденном при спаривании. Всякая группа с некоторым кодом *a* в переменной X (идущей после WITH) меняет свой код на код *b*, если в переменной I (что перед WITH) той группе взаимосоответствует группа с кодом *b*. Другими словами, перекодировка прямо следует табличке «Соответствия», показываемой в результатах. Выберите:

NO - (тж по умолчанию/незаданию) не предпринимать перекодировку.

YES - сделать перекодировку (переменных VARS, стоящих за WITH).

*имя файла* - перекодировку не делать, но написать синтаксис ее. Укажите (в кавычках или апострофах) путь/имя SPS-файла для сохранения. Вы можете, используя синтаксис, сделать перекодировку позже или перекодировать им другой массив. (Если перекодировка фактически не понадобилась, то синтаксис будет содержать перекодировку значений сами в себя.)

RECODE= YES или *имя файла* допустимо только при наличии слова WITH в VARS.

**OTHER**

Данная подкоманда играет роль при RECODE= YES или ‘файл’. Она определяет, как поступить при перекодировании с теми группами переменной X, которым нет в пару соответствующих групп в переменной I.

INDIVID - (тж по умолчанию/незаданию) оставить эти группы индивидуальными.

*значение* - объединить эти группы в одну. Укажите код для этой группы. Требования к допустимым значениям см. в п/к VARS.

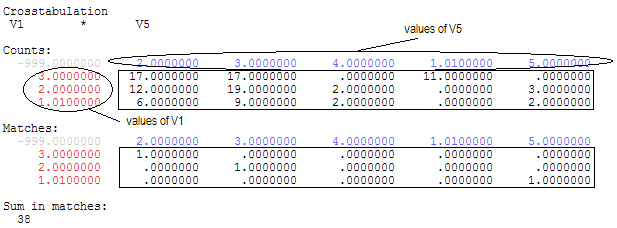
Когда OTHER=INDIVID, макрос сохраняет индивидуальность неспаренных групп так. Если при перекодировке на код такой группы не претендует никакая другая группа, то группа сохраняет свой код. Если же на ее код претендует какая-то другая группа, то код изменяется прибавлением к нему 0.01 повторительно до тех пор, пока он не станет уникальным, т.е. на него не станет других претендентов.

При METHOD=PGREEDY спаривания, которые не были засчитаны, поскольку не удовлетворили условию-ограничению (см.), рассматриваются тоже как группы переменной X, которым нет в пару соответствующих групп в переменной I, поэтому п/к OTHER относится к ним так же.

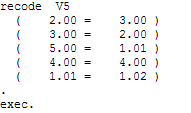
ПРИМЕР 3.

!KO\_grmatch vars= v1 WITH v5 /method= HUNGAR /recode= 'd:\exercise\RecSyntax.sps'.

* Разбиение V1 сопоставляется с разбиением V5. Поскольку есть слово WITH, можно выписать синтаксис для перекодировки групп в V5 в кодировку групп, какая в V1. В окно результатов макрос выдал:



* Кросстабуляция частот (Counts), в которой значения двух переменных – коды групп V1 это ряды, а коды групп V5 это столбцы. Значения макрос не сортирует по возрастанию, они показаны в том порядке, в каком встречены в данных. Мы видим, что в V1 есть три группы с кодами 1.01, 2, 3, а в V5 есть пять групп с кодами 1.01, 2, 3, 4, 5.
* Задача макроса – спарить группы между двумя разбиениями на базе этой таблицы частот. Основанием для спаривания является высокая частота в ячейке таблицы (спаривание старается максимизировать сумму в парах, Sum in Matches). Результаты спаривания показаны в таблице Matches (Соответствия). Группа 2 (V5) спарена (т.е. признана взаимосоответствующей) с группой 3 (V1), группа 3 (V5) спарена с группой 2 (V1), группа 5 (V5) спарена с группой 1.01 (V1). Две группы V5 (4 и 1.01) не спарены ни с одной из групп V1 (спарить группу одной переменной можно не более чем с одной группой другой переменной: взаимно-однозначное соответствие). Макрос выписал синтаксис для перекодировки V5, следуя этим результатам спаривания:



* Коды 2, 3, 5 в V5 должны быть заменены на им отвечающие коды V1. Коды 4 и 1.01 могут быть сохранены, т.к. им нет соответствий в V1. Однако код 1.01 уже занят: этот код существует в V1 и в него будет перекодирован код 5 в V5. Поэтому макрос прибавил 0.01 к коду 1.01 V5, чтобы сохранить его индивидуальность (по умолчанию, OTHER=INDIVID), так что 1.01 предлагается заменить на 1.02. Если бы 1.02 оказался тоже занят, макрос предложил бы 1.03, и т.д.

**SIMMX**

Поскольку при отсутствии слова WITH в VARS сопоставляются между собой попарно все переменные VARS, результаты можно собрать вместе в квадратную симметричную матрицу, рядами и столбцами которой служат все группы всех переменных. В этой матрице взаимосоответствующие пары помечены 1, а остальные пары – 0. Кроме того, для удобства, группы одного и того же разбиения (переменной) имеют на пересечении друг с другом значение минус 1, что подчеркивает факт их альтернативности, эти группы по определению не могут быть соответственными. Таким образом, это матрица сходств с тремя возможными значениями в ней.

Задайте SIMMX=YES, если хотите получить эту матрицу взаимосоответствий. Макрос выведет ее в новый безымянный массив и проделает с ней иерархический кластерный анализ методом полной связи (дальнего соседа). Вы увидите дендрограмму, на которой сразу заметны взаимосоответствующие группы от разных переменных. SIMMX=YES допустимо только при отсутствии слова WITH в VARS. По умолчанию и при SIMMX=NO матрица не выдается и дендрограмма не показывается.

**PRINT**

По умолчанию и при PRINT=YES макрос отчитывается о взаимосоответствиях групп в парах переменных. Если хотите подавить выдачу этих результатах, скомандуйте PRINT=NO; вы должны тогда задать RECODE или SIMMX.

***Особые режимы***

Макрос игнорирует взвешивание (впрочем, он не берет в процедуру наблюдения с пропущенными и неположительными весами) и не рассчитан на расщепленное состояние массива данных. Макрос слушается команд, выбирающих наблюдения (FILTER, USE, SELECT IF, N OF CASES) и преобразований, в том числе под командой TEMPORARY.

1. **Схема решения** асимметричных мер согласия в сравнении кластерных разбиений (т.е. знание меток групп не используется). Индексы могут быть симметричны или асимметричны математически и разбиения могут быть симметричны или асимметричны позиционно. Позиционно симметричные разбиения это когда мы просто сравниваем два альтернативных разбиения. Позиционно асимметричные разбиения это когда одно из двух – "эталонное" и другое – "экспериментное", – разные роли.

   Если индекс симметричный и роли симметричные, то нет проблемы. Если индекс симметричный и роли асимметричные, мы говорим: данный симметричный индекс обслуживает асимметричные роли (как, положим, коэффициент корреляции все еще валидный инлекс в отношениях причина-следствие).

   Если индекс асимметричный и роли симметричные, тогда мы третируем асимметрию индекса как просто неудобство и усредняем/сочетаем два значения (если это действие защитимо).

   Если индекс асимметричный и роли асимметричные, мы можем принять и интерпретировать оба различных значения индекса (плюс, возможно, усредняем/сочетаем их в одно). [↑](#footnote-ref-1)
2. В статье-источнике нижний индекс дан как вместо , но эффект тот же. [↑](#footnote-ref-2)
3. **Схема решения** асимметричных мер согласия в сравнении классификационных разбиений (т.е. знание меток групп используется). Индексы могут быть симметричны или асимметричны математически и разбиения могут быть симметричны или асимметричны позиционно. Позиционно симметричные разбиения это когда мы просто сравниваем два альтернативных разбиения. Позиционно асимметричные разбиения это когда одно из двух – "эталонное" и другое – "экспериментное", – разные роли.

   Если индекс симметричный и роли симметричные, то нет проблемы. Если индекс симметричный и роли асимметричные, мы говорим: данный симметричный индекс обслуживает асимметричные роли (как, положим, коэффициент корреляции все еще валидный инлекс в отношениях причина-следствие).

   Если индекс асимметричный и роли симметричные, тогда мы отказываемся от симметрии ролей и переходим в ситуацию, где одно разбиение объявляется "эталонным" и другое "экспериментным".

   Если индекс асимметричный и роли асимметричные, тогда мы интерпретируем только одно из двух значений индекса, исходя из того, "кто есть кто" в паре разбиений. [↑](#footnote-ref-3)