Comparing Clusterings - An Overview [*]

Silke Wagner Dorothea Wagner

January 12, 2007

1 Introduction

As the amount of data we nowadays have to deal with becomes larger and larger, the methods that help us to detect structures in the data and to identify interesting subsets in the data become more and more important. One of these methods is clustering, i.e. segmenting a set of elements into subsets such that the elements in each subset are somehow "similiar" to each other and elements of different subsets are "unsimilar". In the literature we can find a large variety of clustering algorithms, each having certain advantages but also certain drawbacks. Typical questions that arise in this context comprise:

- Is the algorithm sensitive to small perturbations, i.e. can small changes in the data (so-called "noise") entail large changes in the clustering?
- Is the algorithm sensitive to the order of the data, i.e. can another order of the data result in a very different clustering?
- How similar are the solutions of two different algorithms?
- If an optimal solution is available: How close is the clustering solution to the optimal one?

For examining these aspects, it would be desirable to have a "measure" for the similarity between two clusterings or for their distance[1]. In a more general context, it can be necessary to combine different clusterings to a single one, i.e. calculating a "mean value" of the clusterings. Possible applications are:

[*] The authors gratefully acknowledge financial support from the European Commission within DELIS (contract no. 001907).

[1] Every similarity measure can be transformed into a distance measure and vice versa.

- Combining the results of different algorithms in order to obtain "robust" clusterings [17].

- Intregration of already existing clusterings that have been built before but that cannot be reconstructed (e.g. because the algorithms or features that were used to build them are not known)[16].

- Many companies store their data not only in one database but the data is geographically distributed. Often, it is unfeasible to transfer all data to one place for performing data analysis there because of the high computational, bandwidth and storage costs. Thus, it is desirable to have methods for combining decentrally performed clusterings to one clustering that represents the whole data.

- Legal restrictions force the companies to have several copies of their data, each copy with a different feature set (certain features must not be stored together). For cluster analysis, they have to perform feature distributed clustering and afterwards join the clusterings into one "mean value" clustering.

- In social sciences often arise clustering problems with multiple optimization criteria: a typical example is the "second world war politicians" problem [19], in which many persons were asked to rate the dissimilarities of second world war politicians. Each person corresponds to an optimization criterion. A good clustering of the politicians should be as close as possible to all the personal ratings. A common approach to these multiple criteria clustering problems is the calculation of a "mean value" of the single criterion clusterings.

These are only some applications in which a "mean value" of multiple clusterings is needed. For this purpose we need a distance (or similarity) measure for clusterings. This paper gives an overview and some analysis of the measures that we find in the literature.
In the second section we introduce the basic definitions and some notations. In section three, four and five we present the measures for comparing clusterings that have been presented in the literature so far. The subdivision into three sections corresponds to two "natural" divisions of the measures: Even though the measures can all be derived from the confusion matrix, they base on different ideas: counting of pairs of elements (Section 3), summation of set overlaps (Section 4), and the use of the information-theoretical mutual information (Section 5).

However, the division also reflects the chronological development of the measures: the counting pairs measures date from the 1970s and 1980s, the measures based on set overlaps from the 1990s and the information-theoretical measures have been developed in the last years (2002/2003).
Till this day, there is no formalization of the problem of comparing clusterings. We think that a set of axioms would be helpful in detecting and defining "good" measures. As a first step towards such a set of axioms, we give in Section 6 aspects and properties that have to be taken into account.

2 Definitions and notations

Let X be a finite set with cardinality $|X| = n$. A *clustering* $\mathcal{C}$ is a set $\{C_1, \dots, C_k\}$ of non-empty disjoint subsets of X such that their union equals X. The set of all clusterings of X is denoted by $\mathcal{P}(X)$. For a clustering $\mathcal{C} = \{C_1, \dots, C_k\}$ we assume $|C_i| > 0$ for all $i = 1, \dots, k$. A *trivial clustering* is either the one-clustering that consist of just one cluster or the singleton clustering in which every element forms its own cluster.
Let $\mathcal{C}' = \{C'_1, \dots, C'_\ell\} \in \mathcal{P}(X)$ denote a second clustering of X. The *confusion matrix* $M = (m_{ij})$ (or *contingency table*) of the pair $\mathcal{C}, \mathcal{C}'$ is a $k \times \ell$-matrix whose ij-th entry equals the number of elements in the intersection of the clusters C_i and C'_j:

$$m_{ij} = |C_i \cap C'_j|,\ 1 \leq i \leq k,\ 1 \leq j \leq \ell.$$

Clustering $\mathcal{C}'$ is a *refinement* of $\mathcal{C}$ (and $\mathcal{C}$ is a *coarsening* of $\mathcal{C}'$), if each cluster of $\mathcal{C}'$ is contained in a cluster of $\mathcal{C}$, formally:

$$\forall C'_j \in \mathcal{C}'\ \exists C_i \in \mathcal{C} :\ C'_j \subseteq C_i.$$

The *product* $\mathcal{C} \times \mathcal{C}'$ *of two clusterings* $\mathcal{C}$, $\mathcal{C}'$ is the coarsest common refinement of the two clusterings:

$$\mathcal{C} \times \mathcal{C}' = \{C_i \cap C'_j \mid C_i \in \mathcal{C},\ C'_j \in \mathcal{C}',\ C_i \cap C'_j \neq \emptyset\}.$$

The product $\mathcal{C} \times \mathcal{C}'$ is again a clustering, and if $\mathcal{C}'$ is a refinement of $\mathcal{C}$, then $\mathcal{C} \times \mathcal{C}' = \mathcal{C}'$.

3 Measures based on counting pairs

A very intuitional approach to comparing clusterings is counting pairs of objects that are "classified" in the same way in both clusterings, i.e. pairs of

elements of X that are in the same cluster (in different clusters, respectively) under both clusterings.
The set of all (unordered) pairs of elements of X is the disjoint union of the following sets:

$$\begin{aligned}
S_{11} &= \left\{ \text{ pairs that are in the same cluster under } \mathcal{C} \text{ and } \mathcal{C}\text{' } \right\} \\
S_{00} &= \left\{ \text{ pairs that are in different clusters under } \mathcal{C} \text{ and } \mathcal{C}\text{' } \right\} \\
S_{10} &= \left\{ \begin{array}{l} \text{pairs that are in the same cluster under } \mathcal{C} \text{ but in} \\ \text{different ones under } \mathcal{C}\text{'} \end{array} \right\} \\
S_{01} &= \left\{ \begin{array}{l} \text{pairs that are in different clusters under } \mathcal{C} \text{ but} \\ \text{in the same under } \mathcal{C}\text{'} \end{array} \right\}
\end{aligned}$$

Let $n_{ab} := |S_{ab}|,\ a, b \in \{0, 1\}$, denote the respective sizes. We have

$$n_{11} + n_{00} + n_{10} + n_{01} = \binom{n}{2}.$$

3.1 Chi Squared Coefficient

The most ancient measures for comparing clusterings were originally developed for statistical issues. We want to assert the *Chi Squared Coefficient* as a representative, since it is one of the most well-known measures of this kind. It is defined as

$$\chi(\mathcal{C}, \mathcal{C}') = \sum_{i=1}^{k} \sum_{j=1}^{\ell} \frac{(m_{ij} - E_{ij})^2}{E_{ij}} \quad \text{where } E_{ij} = \frac{|C_i||C'_j|}{n}.$$

As can be seen in [20], there are several variations of the measure. Originally, it was suggested in 1900 by Pearson for testing independance in a bivariate distribution, not for evaluating association (which, in the context of clustering, corresponds to evaluating similarity). The problem in transferring such a measure to the purpose of comparing clusterings lies in the fact that we have to assume independance of the two clusterings. In general, this is not true and therefore the result of a comparison with such a measure has to be challenged (see Sect. 3.6).

3.2 Rand Index

3.2.1 General Rand Index

Rand's Index [1] was motivated by standard classification problems in which the result of a classification scheme has to be compared to a correct classification. The most common performance measure for this problem calculates the fraction of correctly classified (respectively misclassified) elements to all elements. For Rand, comparing two clusterings was just a natural extension of this problem which has a corresponding extension of the performance measure: instead of counting single elements he counts correctly classified pairs of elements. Thus, the Rand Index is defined by:

$$\mathcal{R}(\mathcal{C},\mathcal{C}') = \frac{2(n_{11}+n_{00})}{n(n-1)}$$

$\mathcal{R}$ ranges from 0 (no pair classified in the same way under both clusterings) to 1 (identical clusterings). The value of $\mathcal{R}$ depends on both, the number of clusters and the number of elements. Morey and Agresti showed that the Rand Index is highly dependent upon the number of clusters [2]. In [4], Fowlkes and Mallows show that in the (unrealistic) case of independant clusterings the Rand Index converges to 1 as the number of clusters increases which is undesirable for a similarity measure.

3.2.2 Adjusted Rand Index

The expected value of the Rand Index of two random partitions does not take a constant value (e.g. zero). Thus, Hubert and Arabie proposed an adjustment [3] which assumes a generalized hypergeometric distribution as null hypothesis: the two clusterings are drawn randomly with a fixed number of clusters and a fixed number of elements in each cluster (the number of clusters in the two clusterings need not be the same). Then the adjusted Rand Index is the (normalized) difference of the Rand Index and its expected value under the null hypothesis. It is defined as follows [6]:

$$\mathcal{R}_{adj}(\mathcal{C},\mathcal{C}') = \frac{\sum_{i=1}^{k}\sum_{j=1}^{\ell}\binom{m_{ij}}{2} - t_3}{\frac{1}{2}(t_1+t_2)-t_3}$$

$$\text{where } t_1 = \sum_{i=1}^{k}\binom{|C_i|}{2},\ t_2 = \sum_{j=1}^{\ell}\binom{|C'_j|}{2},\ t_3 = \frac{2t_1t_2}{n(n-1)}$$

This index has expected value zero for independant clusterings and maximum value 1 (for identical clusterings). The significance of this measure has to be put into question because of the strong assumptions it makes on the distribution. Meila [7] notes, that some pairs of clusterings may result in negative index values.

3.3 Fowlkes–Mallows Index

Fowlkes and Mallows introduced their index as a measure for comparing hierarchical clusterings[2] [4]. However, it can also be used for flat clusterings since it consists in calculating an index $\mathcal{B}_i$ for each level $i = 2, \dots, n-1$ of the hierarchies in consideration and plotting $\mathcal{B}_i$ against i. The measure $\mathcal{B}_i$ is easily generalized to a measure for clusterings with different numbers of clusters. The generalized Fowlkes–Mallows Index is defined by

$$\mathcal{FM}(\mathcal{C}, \mathcal{C}') = \frac{\sum_{i=1}^{k} \sum_{j=1}^{\ell} m_{ij}^2 - n}{\sqrt{(\sum_i |C_i|^2 - n)(\sum_j |C'_j|^2 - n)}} = \frac{n_{11}}{\sqrt{(n_{11} + n_{10})(n_{11} + n_{01})}}$$

In the context of Information Retrieval this measure can be interpreted as the geometric mean of precision (ratio of the number of retrieved relevant documents to the total number of retrieved documents $= \frac{n_{11}}{n_{11}+n_{10}}$) and recall (ratio of the number of retrieved relevant documents to the total number of relevant documents $= \frac{n_{11}}{n_{11}+n_{01}}$).
Like for the adjusted Rand Index, the "amount" of similarity of two clusterings corresponds to the deviation from the expected value under the null hypothesis of independant clusterings with fixed cluster sizes. Again, the strong assumptions on the distribution make the result hard to interpret. Futhermore, this measure has the undesirable property that for small numbers of clusters, the value is very high, even for independant clusterings (which even achieve the maximum value for small numbers of clusters). Wallace proposed to attenuate this effect by substracting the number of pairs whose match is forced by the cluster overlaps from the number of "good" pairs and from the number of all pairs [9].

3.4 Mirkin Metric

The Mirkin Metric which is also known as *Equivalence Mismatch Distance* [11] is defined by

$$\mathcal{M}(\mathcal{C}, \mathcal{C}') = \sum_{i=1}^{k} |C_i|^2 + \sum_{j=1}^{\ell} |C'_j|^2 - 2 \sum_{i=1}^{k} \sum_{j=1}^{l} m_{ij}^2.$$

It corresponds to the Hamming distance for binary vectors if the set of all pairs of elements is enumerated and a clustering is represented by a

[2] A hierarchical clustering of a set X is a hierarchy of $|X|$ clusterings, with the two trivial clusterings at the top and bottom, respectively, and each level of the hierarchy is a refinement of all the levels above.

binary vector defined on this enumeration. An advantage is the fact that this distance is a metric on $\mathcal{P}(X)$. However, this measure is very sensitive to cluster sizes such that two clusterings that are "at right angles" to each other (i.e. each cluster in one clustering contains the same amount of elements of each of the clusters of the other clustering) are closer to each other than two clusterings for which one is a refinement of the other [11]. The Mirkin Metric is a variation of the Rand Index [7] since it can be rewritten as

$$\mathcal{M}(\mathcal{C},\mathcal{C}') = 2(n_{01} + n_{10}) = n(n-1)(1 - \mathcal{R}(\mathcal{C},\mathcal{C}')).$$

3.5 Other measures

3.5.1 Jaccard Index

The Jaccard Index is very common in geology and ecology, e.g. for measuring the species diversity between two different communities [10]. It is very similar to the Rand Index, however it disregards the pairs of elements that are in different clusters for both clusterings. It is defined as follows:

$$\mathcal{J}(\mathcal{C},\mathcal{C}') = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$$

3.5.2 Partition Difference

The Partition Difference [19] simply counts the pairs of elements that belong to different clusters unter both clusterings:

$$\mathcal{PD}(\mathcal{C},\mathcal{C}') = n_{00}$$

According to [19], this measure is commonly used. In our opinion, it has too many drawbacks and should therefore not be used: the measure wants to express a distance, but it is not a distance in the mathematical sense, since it fulfills neither the identity of indiscernibles-property (you can have $\mathcal{PD}(\mathcal{C},\mathcal{C}') = 0$, but $\mathcal{C} \neq \mathcal{C}'$, e.g. for the trivial one-clustering $\mathcal{C}$ and an arbitrary clustering $\mathcal{C}' \neq \mathcal{C}$), nor the triangle inequality (take two arbitrary, non-trivial clusterings $\mathcal{C} \neq \mathcal{C}'$ and the trivial one-clustering $\mathcal{C}''$, then $\mathcal{PD}(\mathcal{C},\mathcal{C}'') + \mathcal{PD}(\mathcal{C}'',\mathcal{C}') = 0 < \mathcal{PD}(\mathcal{C},\mathcal{C}')$). The measure is sensitive to cluster sizes and the number of clusters and since it is not normalized, the values are hard to interpret (what does a distance of 5 mean, when we do not know the total number of pairs of elements?).

3.6 General remarks

As mentioned before, the measures presented in this section can all be calculated by means of the confusion matrix M (and the cluster sizes); this is either obvious from the formula (e.g. for the Fowlkes-Mallow Index) or can be seen after some transformation (e.g. for the Rand Index, which can be transformed into a variation of the Mirkin Metric, see 3.4).

For different reasons, these measures do not seem to be very appealing. Some of them are sensitive to certain parameters (cluster sizes, number of clusters); think of a pair of clusterings with similarity $\alpha \in [0, 1]$ and replace each element in the underlying set by two elements. Why should the resulting pair of clusterings have a similarity other than α? This behavior, as well as sensitivity to the number of clusters, are undesirable.
Other measures, like the Fowlkes-Mallows Index, suffer from another drawback: they make use of a very strong null hypothesis, that is, independance of the clusterings, fixed number of clusters, and fixed cluster sizes. When comparing results provided by clustering algorithms these assumptions are - apart from the number of clusters that is fixed for some algorithms - violated. None of the algorithms works with fixed cluster sizes. Furthermore, in practice it would be against the intuition to compare two clusterings when assuming that there is no relationship between them. In fact, we compare clusterings because we suppose a certain relationship and we want to know how strong it is [12].

4 Measures based on set overlaps

Another kind of measure tries to match clusters that have a maximum absolute or relative overlap. This is also a quite intuitional approach, however the assymetry of some of the measures makes them difficult to use.

4.1 $\mathcal{F}$-Measure

The $\mathcal{F}$-Measure has its origin in the field of document clustering where it is used to evaluate the accuracy of a clustering solution. Each cluster of the first clustering is a (predefined) class of documents and each cluster of the second clustering is treated as the result of a query [14], [13]. The $\mathcal{F}$-Measure for a cluster C'_j with respect to a certain class C_i indicates how "good" the cluster C'_j describes the class C_i by calculating the harmonic mean of precision $p_{ij} = \frac{m_{ij}}{|C'_j|}$ and recall $r_{ij} = \frac{m_{ij}}{|C_i|}$ for C'_j and C_i:

$$\mathcal{F}(C_i, C'_j) = \frac{2 \cdot r_{ij} \cdot p_{ij}}{r_{ij} + p_{ij}} = \frac{2|C_i||C'_j|}{|C_i| + |C'_j|}$$

The overall $\mathcal{F}$-Measure is then defined as the weighted sum of the maximum $\mathcal{F}$-Measures for the clusters in $\mathcal{C}'$:

$$\mathcal{F}(\mathcal{C}, \mathcal{C}') = \mathcal{F}(\mathcal{C}') = \frac{1}{n} \sum_{i=1}^{k} n_i \max_{j=1}^{\ell} \{\mathcal{F}(C_i, C'_j)\}$$

It can easily be seen that this measure is not symmetric. Thus, this may be an appropriate index for comparing a clustering with an optimal clustering solution. However, in general the optimal solution is not know, which makes an assymetric measure hard to interpret.

In [7], Meila claims, that in [13], Larsen uses a variation of this measure which is normalized by the number of clusters instead of the number of elements. She gives an example where this "Larsen-measure" has a very strange behavior. However, as can be seen in [13], Larsen does not use this measure, but also the $\mathcal{F}$-Measure as defined above. Actually, other authors, as for example Steinbach, Kapyris and Kumar in [15], or Fung in [14], refer to Larsen when introducing the $\mathcal{F}$-Measure.

4.2 Meila-Heckerman- and Maximum-Match-Measure

In [8], Meila and Heckerman use another assymetric measure, which they apply to comparing clustering algorithms. For their study, they do not compare the results of the different clustering methods among each other, but they compare each clustering result with an optimal clustering solution (their study is on synthetic data). For this purpose they use the following measure:

$$\mathcal{MH}(\mathcal{C}, \mathcal{C}') = \frac{1}{n} \sum_{i=1}^{k} \max_{C'_j \in \mathcal{C}} m_{ij}$$

where $\mathcal{C}$ is the clustering that is provided by the algorithm and $\mathcal{C}'$ is the optimal clustering. As for the preceding measure, its assymetry makes it inappropriate for the general task of comparing clusterings. However, it can be generalized to the symmetric *Maximum-Match-Measure* $\mathcal{M}(\mathcal{C}, \mathcal{C}')$ which can be determined as follows: look for the largest entry m_{ab} of the confusion matrix M and match the corresponding clusters C_a and C'_b (this is the cluster pair with the largest (absolute) overlap). Afterwards cross out the a-th row and the b-th column and repeat this step (searching for the maximum entry, matching the corresponding clusters and deleting the corresponding row and column) until the matrix has size 0. Afterwards you sum up the matches and divide it by the total number of elements:

$$\mathcal{MM}(\mathcal{C},\mathcal{C}') = \frac{1}{n} \sum_{i=1}^{\min\{k,\ell\}} m_{ii'}$$

where i' is the index of the cluster in $\mathcal{C}$' that is matched to cluster $C_i \in \mathcal{C}$. Note, that in the case of $k \neq \ell$, this measure completely disregards the $|k-\ell|$ "remaining" clusters in the clustering with the higher cardinality.

4.3 Van Dongen-Measure

In [11], van Dongen introduces a symmetric measure, that is also based on maximum intersections of clusters. It is defined as follows:

$$\mathcal{D}(\mathcal{C},\mathcal{C}') = 2n - \sum_{i=1}^{k} \max_j \ m_{ij} - \sum_{j=1}^{\ell} \max_i \ m_{ij}$$

This measure has the nice property of being a metric on the space of all clusterings of the underlying set X. However, it ignores the parts of the clusters outside the intersections (see 4.4).

4.4 General remarks

The preceding measures have the common property of just taking the overlaps into account. They completely disregard the unmatched parts of the clusters (or even complete clusters, as the Maximum-Match-Measure). In [7], Meila presents a nice example that points out the negative effect of this "behavior" of a measure: take a clustering $\mathcal{C}$ with k equal clusters and derive two variations $\mathcal{C}'$ and C'' as follows: $\mathcal{C}'$ is obtained from $\mathcal{C}$ by shifting a fraction α of the elements in each cluster C_i to the "next" cluster $C_{i+1 \mod k}$. The clustering $\mathcal{C}''$ is obtained from $\mathcal{C}$ by reassigning a fraction α of the elements in each cluster C_i evenly between the other clusters. If $\alpha < 0.5$, then $\mathcal{F}(\mathcal{C},\mathcal{C}') = \mathcal{F}(\mathcal{C},\mathcal{C}'')$, $\mathcal{MH}(\mathcal{C},\mathcal{C}') = \mathcal{MH}(\mathcal{C},\mathcal{C}'')$, $\mathcal{MM}(\mathcal{C},\mathcal{C}') = \mathcal{MM}(\mathcal{C},\mathcal{C}'')$ and $\mathcal{D}(\mathcal{C},\mathcal{C}') = \mathcal{D}(\mathcal{C},\mathcal{C}'')$, which means that for all the measures $\mathcal{C}''$ is as similar to $\mathcal{C}$ as $\mathcal{C}'$. This contradicts our intuition that $\mathcal{C}'$ is a less modified version of $\mathcal{C}$ than $\mathcal{C}''$ and is therefore not desirable.

5 Measures based on Mutual Information

This approach to the comparison of clusterings has its origin in information theory and is based on the notion of *entropy*:
The entropy S for an information, e.g. a text T, with alphabet Σ is defined as

$$S(T) = -\sum_{i \in \Sigma} p_i \log_2(p_i)$$

where p_i is the probability of finding i in T (more precisely, we have a discrete random variable Y taking $|\Sigma|$ values and $P(Y = i) = p_i$). The entropy is measured in bits and $S(T) \cdot |T|$ is the number of bits that is needed for representing T [18].
When applied to clusterings, the meaning of entropy can be described as follows [7]: assuming that all elements of X have the same probability of being picked and choosing an element of X at random, the probability that this element is in cluster $C_i \in \mathcal{C}$ is $P(i) = \frac{|C_i|}{n}$. Then, the *entropy associated with clustering* $\mathcal{C}$ is

$$\mathcal{H}(\mathcal{C}) = -\sum_{i=1}^{k} P(i) \log_2 P(i).$$

Informally, the entropy of a clustering $\mathcal{C}$ is a measure for the uncertainty about the cluster of a randomly picked element. In the case of a trivial clustering (one cluster or n clusters), we know the cluster of a randomly picked element, thus the entropy of such a clustering is 0.

The notion of entropy can be extended to that of *mutual information*, which describes how much we can on the average reduce the uncertainty about the cluster of a random element when knowing its cluster in another clustering of the same set of elements. Formally, the *mutual information between two clusterings* $\mathcal{C}$, $\mathcal{C}'$ is defined as

$$\mathcal{I}(\mathcal{C}, \mathcal{C}') = \sum_{i=1}^{k} \sum_{j=1}^{\ell} P(i,j) \log_2 \frac{P(i,j)}{P(i)P(j)}$$

where $P(i,j)$ is the probability that an element belongs to cluster C_i in $\mathcal{C}$ and to cluster C'_j in $\mathcal{C}'$:

$$P(i,j) = \frac{|C_i \cap C'_j|}{n}.$$

The mutual information $\mathcal{I}$ is a metric on the space of all clusterings. However, it is not bounded by a constant value which makes it difficult to interpret. As the mutual information between two clusterings is bounded by their entropies,

$$\mathcal{I}(\mathcal{C}, \mathcal{C}') \leq \min\{\mathcal{H}(\mathcal{C}), \mathcal{H}(\mathcal{C}')\},$$

a normalization by the geometric or arithmetic mean of the entropies seems to be reasonable. These normalizations were proposed by Strehl & Gosh and Fred & Jain, respectively.

5.1 Normalized Mutual Information by Strehl & Ghosh

In [16], Strehl and Ghosh introduce the problem of combining multiple clusterings into a single one without accessing the original features or algorithms that determined these clusterings. For this purpose, they (approximately) determine the clustering that has the maximal average normalized mutual information with all the clusterings in consideration, where the *normalized mutual information* between two clusterings is defined as

$$\mathcal{NMI}_1(\mathcal{C}, \mathcal{C}') = \frac{\mathcal{I}(\mathcal{C}, \mathcal{C}')}{\sqrt{\mathcal{H}(\mathcal{C})\mathcal{H}(\mathcal{C}')}}.$$

Since $\mathcal{H}(\mathcal{C}) = \mathcal{I}(\mathcal{C}, \mathcal{C})$, Strehl & Gosh prefer the normalization by the geometric mean because of the analogy with a normalized inner product in a Hilbert space. We have

$$0 \leq \mathcal{NMI}_1(\mathcal{C}, \mathcal{C}') \leq 1$$

with $\mathcal{NMI}_1(\mathcal{C}, \mathcal{C}') = 1$ for $\mathcal{C} = \mathcal{C}'$ and $\mathcal{NMI}_1(\mathcal{C}, \mathcal{C}') = 0$ if for all i, $1 \leq i \leq k$, and for all j, $1 \leq j \leq \ell$, we have $P(i,j) = 0$ or $P(i,j) = P(i) \cdot P(j)$. The treatment of the special case when one of the clusterings is trivial (e.g. the denominator of the fraction becomes 0) is not mentioned in [16].

5.2 Normalized Mutual Information by Fred & Jain

In order to obtain a good and robust clustering of a given set of elements, Fred and Jain propose to combine the results of multiple clusterings instead of using just one particular algorithm [17]. Ideally, the solution should satisfy three properties:

1. Consistency with the set of clusterings.

2. Robustness to small variations in the set of clusterings.

3. Goodness of fit with the ground truth information (if available).

Property 1 is modelled by an optimization criterion searching for the clustering that maximizes the average normalized mutual information with all the clusterings, where the *normalized mutual information* between two clusterings is defined as

$$\mathcal{NMI}_2(\mathcal{C},\mathcal{C}') = \frac{2\mathcal{I}(\mathcal{C},\mathcal{C}')}{\mathcal{H}(\mathcal{C}) + \mathcal{H}(\mathcal{C}')}.$$

For details about the modelling of property 2, see [17] (property 3 is not included in the model but used for the evaluation of the results).
As for the previous index, we have

$$0 \le \mathcal{NMI}_2(\mathcal{C},\mathcal{C}') \le 1$$

with $\mathcal{NMI}_2(\mathcal{C},\mathcal{C}') = 1$ for $\mathcal{C} = \mathcal{C}'$ and $\mathcal{NMI}_2(\mathcal{C},\mathcal{C}') = 0$ if $P(i,j) = 0$ or $P(i,j) = P(i) \cdot P(j)$ for all i, $1 \le i \le k$, and for all j, $1 \le j \le \ell$. The case of denominator 0 (both clusterings are trivial) is not mentioned.

5.3 Variation of Information

Meila proposes in [7] another measure based on the entropy, which is called *variation of information* between two clusterings (by analogy to the total variation of a function):

$$\begin{aligned} \mathcal{VI}(\mathcal{C},\mathcal{C}') &= \mathcal{H}(\mathcal{C}) + \mathcal{H}(\mathcal{C}') - 2\mathcal{I}(\mathcal{C},\mathcal{C}') \\ &= [\mathcal{H}(\mathcal{C}) - \mathcal{I}(\mathcal{C},\mathcal{C}')] + [\mathcal{H}(\mathcal{C}') - \mathcal{I}(\mathcal{C},\mathcal{C}')] \qquad (1) \end{aligned}$$

Informally, the first term of (1) corresponds to the amount of information about $\mathcal{C}$ that we loose, while the second term corresponds to the amount of information about $\mathcal{C}'$ that we still have to gain, when going from clustering $\mathcal{C}$ to $\mathcal{C}'$.
The variation of information is the only information-theoretical measure of which a more detailed analysis can be found in the literature. In the following, we will summarize the main properties of $\mathcal{VI}$, for proofs and details see [7].

- $\mathcal{VI}(\mathcal{C},\mathcal{C}')$ is a metric on $\mathcal{P}(X)$.
- $\mathcal{VI}(\mathcal{C},\mathcal{C}')$ is not bounded by a constant value. However, there is an upper bound of $\log n$ (which is attained for all n, e.g. with the two trivial clusterings) and if the number of clusters is bounded by a constant K

with $K \leq \sqrt{n}$, then $\mathcal{VI}(\mathcal{C}, \mathcal{C}') \leq 2 \log K$; this bound is attained if n is a multiple of K^2.
This means that for large enough n, clusterings of different data sets, with different numbers of elements, but with bounded numbers of clusters are on the same scale in the metric $\mathcal{VI}$. This allows us to compare, add or substract $\mathcal{VI}$-distances across different clustering spaces independently of the underlying data set.

- The product of two clusterings $\mathcal{C}, \mathcal{C}'$ is "collinear" with the two clusterings:
$$\mathcal{VI}(\mathcal{C}, \mathcal{C}') = \mathcal{VI}(\mathcal{C}, \mathcal{C} \times \mathcal{C}') + \mathcal{VI}(\mathcal{C} \times \mathcal{C}', \mathcal{C}').$$
This also implies $\mathcal{VI}(\mathcal{C}, \mathcal{C}') \geq \mathcal{VI}(\mathcal{C}, \mathcal{C} \times \mathcal{C}')$ with equality for $\mathcal{C}' = \mathcal{C} \times \mathcal{C}'$. Thus, the nearest neighbor of a clustering $\mathcal{C}$ is either a refinement of $\mathcal{C}$ or a clustering whose refinement is $\mathcal{C}$. In fact, the nearest neighbor of a clustering is obtained by splitting one element off the smallest cluster (or by the corresponding merge process). This means that small changes in a clustering result in small $\mathcal{VI}$-distances.

- The $\mathcal{VI}$-distance between two clusterings $\mathcal{C}, \mathcal{C}'$ with $\mathcal{C} \neq \mathcal{C}'$ has a lower bound of $\frac{2}{n}$. Thus, with increasing n, the space of clusterings gets a finer granularity.

- $\mathcal{VI}(\mathcal{C}, \mathcal{C}')$ can be computed in $\mathcal{O}(n + k\ell)$: we need time $\mathcal{O}(n)$ for computing the confusion matrix M and time $\mathcal{O}(k\ell)$ for computing $\mathcal{VI}(\mathcal{C}, \mathcal{C}')$ from M.

In [7], Meila gives extensions of the variation of information to soft clusterings (each element belongs with a certain probability to each of the clusters) and to weighted elements (each element is weighted by a non-uniform probability).

5.4 General remarks

The measures based on information-theoretical considerations seem to be quite promising because they do not suffer from the drawbacks that we can find for measures that are based on counting pairs or on set overlaps. However, they possibly suffer from other disadvantages that we do not know yet. Here, an extensive examination of these measures, especially the two versions of the normalized mutual information, is necessary.

6 A first step towards a formalization

We have tried to give an overview of the measures for comparing clusterings that can be found in the literature. The authors of the respective books or papers had different motivations for looking for such a measure, however, none of them tried to formalize the claims he makes on the measure. We assume that they all have certain "basic" claims in common, as e.g. symmetry (asymmetric measures like the Meila-Heckerman-Measure are used for the special case of comparing a clustering with an optimal solution), which should be captured in a set of axioms. As a first step towards a complete set of axioms defining a "good" measure for comparing clusterings, we want to mention properties and aspects that have to be taken into account for this purpose. Let f be a measure for comparing clusterings of a set X and let $\mathcal{C}$, $\mathcal{C}'$, $\mathcal{C}'' \in \mathcal{P}(X)$ be clusterings of X. f should have the following properties:

1. Metric on $\mathcal{P}(X)$, i.e. we have

   (a) Positivity: $f(\mathcal{C},\mathcal{C}') \geq 0$

   (b) $f(\mathcal{C},\mathcal{C}) = 0$

   (c) Symmetry: $f(\mathcal{C},\mathcal{C}') = f(\mathcal{C}',\mathcal{C})$

   (d) Identity of indiscernibles: $f(\mathcal{C},\mathcal{C}') = 0 \Rightarrow \mathcal{C} = \mathcal{C}'$

   (e) Triangle inequality: $f(\mathcal{C},\mathcal{C}'') \leq f(\mathcal{C},\mathcal{C}') + f(\mathcal{C}',\mathcal{C}'')$

   A less restrictive version, claiming only a distance measure (or semimetric) is imaginable, either (which means that we would pass on the identity of indiscernibles and the triangle inequality).
   We prefer the formulation in which the value of f for two clusterings represents a distance, but it can as well be formulated with f representing a similarity. In the literature, there are different definitions of similarity measures, which mostly differ only in the range of the function: often, it is the unit interval (as for example in [21]), sometimes it is an arbitrary interval $[a, b] \subset \mathbb{R}$ (as for example in [22]). Thus, when following the majority and taking the unit interval as range of the function, we can express this property in terms of similarity as follows:
   f is a metrical similarity measure on $\mathcal{P}(X)$, i.e. we have

   (a) $0 \leq f(\mathcal{C},\mathcal{C}') \leq 1$

   (b) $f(\mathcal{C},\mathcal{C}) = 1$

   (c) $f(\mathcal{C},\mathcal{C}') = f(\mathcal{C}',\mathcal{C})$

   (d) $f(\mathcal{C},\mathcal{C}') = 1 \Rightarrow \mathcal{C} = \mathcal{C}'$

   (e) $|f(\mathcal{C},\mathcal{C}') + f(\mathcal{C}',\mathcal{C}'')|f(\mathcal{C},\mathcal{C}'') \leq f(\mathcal{C},\mathcal{C}')f(\mathcal{C}',\mathcal{C}'')$

   Analogous to a metric, a less restrictive version is possible, claiming only the first three properties, which means that f is a (non-metrical) similarity measure. Note, that property (e) corresponds to the triangle inequality for distance measures.

2. No additional constraints, neither on the structure of the clusterings nor on their relation, i.e. no assumptions on

   (a) the cluster sizes

   (b) the number of clusters (particularly, $|\mathcal{C}|$ and $|\mathcal{C}'|$ need not be the same)

   (c) dependencies between $\mathcal{C}$ and $\mathcal{C}'$.

   This property assures, that the result cannot be adulterated by assumptions on the clusterings which, in general, are not fulfilled. As we have seen in Sect. 3, there are measures like the Adjusted Rand Index or the Fowlkes–Mallows Index, that use the expected value under a null hypothesis, assuming independent clusterings and fixed cluster sizes. In general, both assumptions are violated, since none of the well-established clustering algorithms works with fixed cluster sizes and since both clusterings are obtained by clustering the same data set, which means that there is a relationship between the two clusterings (and by applying the distance or similarity measure we want to find out how strong it is). Thus, measures that make such assumptions on the structure of the clusterings and their relation, cannot yield reliable results.

3. Independence from the number of clusters.

   Measures like the General Rand Index (Sect. 3.2.1) have the undesirable property of being highly dependant upon the number of clusters: random clusterings can have a high similarity just because they have a large number of clusters. In order to avoid this we need independence from the number of clusters.

4. Independence from the number of elements

   With this property, we make sure that clusterings of different data sets with different numbers of elements are on the same scale. This is important for comparing distance values of clustering pairs of different data sets. Often, we do this kind of comparison "automatically", e.g. by saying that two clusterings that have distance 0.1 are more similar than two clusterings of another data set that have distance 0.3. However, this conclusion is only true if the values are on the same scale. When we want to compare clustering algorithms, this property allows us to calculate the "mean error" for different algorithms, which is the average of the distances to the optimal clustering for different data sets.

When thinking of desirable properties of a measure for comparing clusterings, a lot of other properties are imaginable. However, they arise rather from special applications than from the general case and go beyond the "basics" for the general case. An example for such a property would be the possibility of extending the domain of the function f to the set of all clusterings of two different underlying sets. Of course, such a comparison only makes sense when the two underlying sets have a non-empty intersection. This property is very important for dynamic clustering, that is, updating a clustering when the underlying set changes (and thus avoiding the reclustering of the whole data set). In the context of dynamic clustering we want to answer questions like:

- How large is the distance between the updated clustering and the reclustering of the updated data set?
- How "far" is the updated clustering from the optimal one (if it is known)?
- Can small changes in the data set cause large changes in the clustering?

Note, that for dynamic clustering "changes in the data" has another meaning than for static clustering: In the static case, small changes in the data mean noise (small changes in the values of attributes of the elements) whereas in dynamic clustering, it means a small number of update operations (insertion or deletion of elements or changes in the similarity between the elements). Thus, in the first case we still have a one-to-one correspondence of the elements in the two data sets while in the second case we loose it.
The first two questions reduce to the "standard" task of comparing clustering. However, for answering the third question, we have to compare clusterings of two different underlying data sets and therefor we need a measure that can be extended appropriately.

References

[1] Rand, William M.: Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association, 66(336):846–850, 1971.

[2] Morey, L. C., Agresti, A.: An Adjustment to the Rand Statistic for Chance Agreement. The Classification Society Bulletin 5, 9-10.

[3] Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification, 2:193–218, 1985.

[4] Fowlkes, E. B., Mallows, C. L.: A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association, 78(383):553–569, 1983.

[5] Ruzzo, Walter L., Yeung, Ka Yee: Details of the Adjusted Rand index and Clustering algorithms.

[6] Kuncheva, Ludmila I., Hadjitodorov, Stefan T.: Using Diversity in Cluster Ensembles. IEEE SMC International Conference on Systems, Man and Cybernetics, 2004.

[7] Meila, Marina: Comparing Clusterings. COLT 2003.

[8] Meila, Marina, Heckerman, David: An Experimental Comparison of Model-based Clustering Methods. Proceedings of the Conference on Knowledge Discovery and Data Mining, pages 16-22, 1999.

[9] Wallace, David L.: A Method for Comparing Two Hierarchical Clusterings: Comment. Journal of the American Statistical Association, 78(383):569–576, 1983.

[10] Hultsch, Linda: Untersuchung zur Besiedlung einer Sprengfläche im Pockautal durch die Tiergruppen Heteroptera (Wanzen) und Auchenorrhyncha (Zikaden). http://www.goek.tu-freiberg.de/oberseminar/OS_03_04/Linda_Hultsch.pdf

[11] van Dongen, Stijn: Performance Criteria for Graph Clustering and Markov Cluster Experiments. Technical Report INS–R0012, Centrum voor Wiskunde en Informatica, 2000.

[12] Jolliffe, Ian T., Morgan, Byron J.T.: A Method for Comparing Two Hierarchical Clusterings: Comment. Journal of the American Statistical Association, 78(383):580–581, 1983.

[13] Larsen, B., Aone, C.: Fast and Effective Text Mining Using Linear Time Document Clustering. Proceedings of the KDD, 16-29, 1999.

[14] Fung, Benjamin C.H.: Hierarchical Document Clustering Using Frequent Itemsets. Masters Thesis, Simon Fraser University, 2002.

[15] Steinbach, M., Karypis, G., Kumar, V.: A Comparison of Document Clustering Techniques. KDD Workshop on Text Mining, 2000.

[16] Strehl, Alexander, Ghosh, Joydeep: Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions. Journal of Machine Learning Research, 3:583–617, 2002.

[17] Fred, Ana L.N., Jain, Anil K.: Robust Data Clustering. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR, (3):128-136, 2003.

[18] Cover, Thomas M., Thomas, Joy A.: Elements of Information Theory. Wiley, 1991.

[19] Li, Tao, Ogihara, Mitsunori, Ma, Sheng: On Combining Multiple Clusterings. Proceedings of the ACM Conference on Information and Knowledge Management, (13):294-303, 2004.

[20] Mirkin, Boris: Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables. The American Statistician, 55, (2): 111-120, 2001.

[21] Bock, Hans Hermann: Automatische Klassifikation. Vandenhoeck und Ruprecht, 1974.

[22] Steinhausen, Detlef, Langer, Klaus: Clusteranalyse – Einführung in die Methoden und Verfahren der automatischen Klassifikation. Walter de Gruyter & Co., 1977.