Precision-recall space to correct external indices for biclustering

Blaise Hanczar BLAISE.HANCZAR@PARISDESCARTES.FR
LIPADE, University Paris Descartes, 45 rue des saint-peres, 75006, Paris, France.

Mohamed Nadif MOHAMED.NADIF@PARISDESCARTES.FR
LIPADE, University Paris Descartes, 45 rue des saint-peres, 75006, Paris, France.

Abstract

Biclustering is a major tool of data mining in many domains and many algorithms have emerged in recent years. All these algorithms aim to obtain coherent biclusters and it is crucial to have a reliable procedure for their validation. We point out the problem of size bias in biclustering evaluation and show how it can lead to wrong conclusions in a comparative study. We present the theoretical corrections for all of the most popular measures in order to remove this bias. We introduce the corrected precision-recall space that combines the advantages of corrected measures, the ease of interpretation and visualization of uncorrected measures. Numerical experiments demonstrate the interest of our approach.

1. Introduction

In many domains more and more data are produced because of recent technological advances and the increasing capabilities of modern computers. However, extracting relevant information from these enormous volumes of data still remains a difficult task. This is why, there has been an increasing interest in the data mining and machine learning methods to handle data. In this paper, we are interested in the problem of biclustering, also called co-clustering (Dhillon, 2001), simultaneous clustering or block clustering (Govaert & Nadif, 2008). All of these algorithms aim to obtain homogeneous or coherent biclusters. The proposed algorithms differ in the patterns they seek, the types of data they are applied to, and the assumptions on which they rest. In recent years biclustering has become an important challenge in data mining and in particular in text mining (Dhillon, 2001) and bioinformatics (Lazzeroni & Owen, 2000; Hanczar & Nadif, 2010; Madeira & Oliveira, 2004; Hanczar & Nadif, 2012).

Several biclustering or co-clustering methods are proposed and all of them claim to be better than the previous ones. A reliable procedure of performance estimation is therefore necessary to evaluate and compare the results of these algorithms. Although this evaluation is crucial in all biclustering publications, few works have focused on this problem (Prelic et al., 2006). The evaluation and comparison of biclustering algorithms is based on performance indices that can be divided into two categories: external and internal indices. The external indices estimate the similarity between a biclustering solution and a priori knowledge. Generally external indices are used to compare a *bicluster solution* produced by a biclustering algorithm with the true biclustering solution. The internal indices compare intrinsic information about data with the biclustering solution produced by an algorithm. In this case, no a priori information further than the raw data is available. Internal indices are not as precise as external indices, but they are important when a priori information is not available (Handl et al., 2005; Lee et al., 2011). Only external indices produce an objective evaluation of the biclustering performance. The internal indices are subjective since they depend on assumptions that correspond more or less to the reality. Unfortunately, in practical applications, the true biclusters are generally unknown and external indices can not be used. A reliable evaluation procedure of a biclustering algorithm should include two steps. The first one consists in testing the biclustering algorithm on artificial datasets where the true biclusters are known. The external indices are used to measure the performance of the tested algorithm and analyze its behavior with different parameters. In the second step the biclustering algorithm is applied to real data and the obtained results are evaluated with internal

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28. Copyright 2013 by the author(s).

indices. If the results on artificial and real data go in the same direction, we can draw reliable conclusions. In the following, we focus on the external indices. It should be notice that although the biclustering evaluation problem has strong connections with the clustering evaluation problem, there are important differences. A bicluster is not just the union of a set of features and a set of examples, we have to consider the structure in two dimensions formed by these sets. Moreover, in the clustering tasks a partition of the elements is computed and evaluated. In biclustering tasks, generally a large part of the points do not belong to any biclusters and some biclusters may overlap, i.e. some points belong to several biclusters. For these reasons, several classic performance measures of clustering can not be used in biclustering. In this paper we present and analyse the main external indices in the precision-recall space. We show that these indices are affected by a size bias that advantages the large biclusters. We give the theoretical correction for each measure in order to remove the size bias. We define the corrected precision-recall space in which the uncorrected measures are not affected by the size bias.

The outline of the paper is as follows. In section 2 we present the main external indices used in biclustering context, we analyze their behavior in the precision-recall space and we show the impact of the tradeoff between precision and recall on the results of these indices. In section 3 we point out the problem of size bias. Section 4 gives the correction to apply to the different measures. In section 5, we introduce the corrected precision-recall space. Section 6 is devoted to present numerical experiments pointing out the advantage to consider the corrected measures and the corrected-space for the biclustering evaluation. Finally, the conclusion summarizes the main contributions.

2. External Measures for Biclustering

Let D be a data matrix where F and E are respectively the set of features and examples. We consider that this matrix contains a "true" biclustering solution corresponding to a set of K biclusters noted $= \{B_1, ..., B_K\}$. A biclustering algorithm, whose the objective is to find out the true biclustering solution , produces an estimated biclustering solution $= \{X_1, ..., X_L\}$.

The external indices consist in evaluating the similarity between the true and estimated biclustering solution. There exist several indices but all of them are based on the following formula:

$$I(\; , \;) = \frac{1}{K} \sum_{i=1}^{K} max_{j=1..L} M(B_i, X_j) \tag{1}$$

Each true bicluster is associated to the estimated bicluster that maximizes the measure M and all values are averaged. The difference between the different indices depends only on the definition of M that measures the similarity between two biclusters.

Let $B = E_B \times F_B$ be a *true* bicluster defined by the set of features F_B and the set of examples E_B and its estimated bicluster $X = E_X \times F_X$, defined by the sets of features F_X and examples E_X. Let's $|D| = |F||E|$, $|B| = |F_B||E_B|$, $|X| = |F_X||E_X|$ and $|B \cap X| = |F_B \cap F_X||E_B \cap E_X|$ be the sizes of the data matrix, true bicluster, estimated bicluster and intersection between true and estimated error respectively. The performance of the estimated bicluster depends on how it matches the true bicluster. Two types of errors can be defined: the points belonging to the true bicluster and not covered by the estimated bicluster, represented by the dark gray area and the points in the estimated bicluster but not belonging to the true bicluster, representing by the light gray area. These two types of error are represented by the notion of precision and recall. The precision is the proportion of points of the estimated bicluster belonging to the true bicluster, it takes this form

$$pre = precision(B, X) = \frac{|B \cap X|}{|X|}. \tag{2}$$

The recall is the proportion of points of the true bicluster covered by the estimated bicluster. It can be written as follows

$$rec = recall(B, X) = \frac{|B \cap X|}{|B|}. \tag{3}$$

We consider here three popular measures used in the biclustering problem: Dice, Jaccard and goodness measures.

2.1. Dice Measure

The Dice measure is the ratio between the intersection and the size of true and estimated biclusters.

$$M_{dice}(B, X) = \frac{2|B \cap X|}{|B| + |X|}. \tag{4}$$

Plugging (2) and (3) in (4), the Dice measure becomes

$$M_{dice}(B, X) = 2\frac{pre.rec}{pre + rec}. \tag{5}$$

The Dice measure is the harmonic mean of precision and recall, it also corresponds to the traditional balanced F-score.

2.2. Jaccard Measure

The Jaccard measure is the ratio between the intersection and union of true and estimated biclusters.

$$M_{jaccard}(B,X) = \frac{|B \cap X|}{|B| + |X| - |B \cap X|}. \tag{6}$$

Plugging (2) and (3) in (6), the Jaccard takes the following form

$$M_{jaccard}(B,X) = \frac{pre.rec}{pre + rec - pre.rec} \tag{7}$$

Note that M_{dice} and $M_{jaccard}$ are compatible i.e. let X_1 and X_2 be two biclusters $M_{dice}(B,X_1) \leq M_{dice}(B,X_2) \iff M_{jaccard}(B,X_1) \leq M_{jaccard}(B,X_2)$.

2.3. Goodness Measure

We also consider a third measure that is the mean of precision and recall. It has not conventional names, it is called here the Goodness measure:

$$M_{good} = \frac{1}{2}(pre + rec). \tag{8}$$

3. Precision-Recall Space

3.1. Definitions and properties

An efficient analysis and visualization tool for these measures is the precision-recall space. It is a 2D space, where the performance of an estimated bicluster is represented by a point in this space (Figure 1).

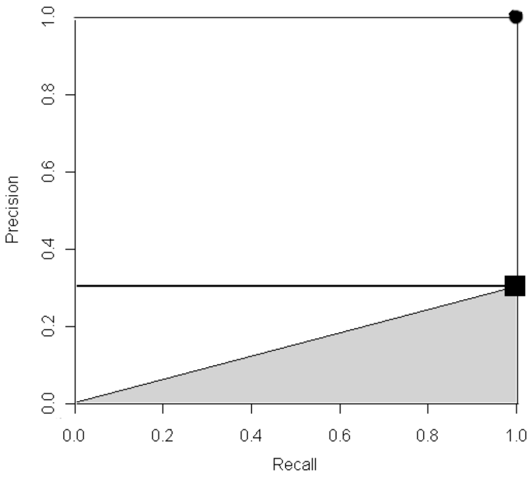

Figure 1. Precision-Recall space

The precision-recall space is close to the ROC space that is defined by the false and true positive rate. Some relationships have been identified between precision-recall and ROC spaces (Davis & Goadrich, 2006). A point on the precision-recall space represents the performance of all biclusters with the same size $|X|$ and the same intersection $|B \cap X|$. The point $(1,1)$ (black dot), maximising both precision and recall, represents the perfect bicluster, i.e. equal to the true bicluster. The point $(1, \frac{|B|}{|D|})$ (black square) represents the case where the estimated bicluster is equal to the whole data matrix $X = D$. The horizontal bold line corresponds to the expected performances of a random bicluster, i.e. bicluster where the set of examples and features are randomly selected. Since it depends on the size of the true bicluster $|B|$, the expected precision of a random bicluster is constant and $[pre] = \frac{|B|}{|D|}$. The expected recall of a random bicluster depends on the size of the estimated bicluster $|X|$, it is equal to $[rec] = \frac{|X|}{|D|}$. The gray area represents performances that cannot be reached by a bicluster. From (2) and (3), note that $pre \geq \frac{|B|}{|D|}rec$ since $|X| \leq |D|$, then all estimated biclusters whose the performance are represented by a point on the line $pre = \frac{|B|}{|D|}rec$ are the biclusters with the minimal intersection possible $|B \cap X|$ for a given size $|X|$. The point (0,0) represents all biclusters whose the intersection with the true bicluster is null.

The behavior of the different measures can be illustrated by the isometrics. The isometrics are collections of points in the precision-recall space with the same value for the metric (Flach, 2003). They are represented in the precision-recall space by lines or curves. The isometrics of precision are represented by horizontal lines and isometrics of recall by vertical lines. The three panels of the figure 2 show respectively the isometrics of Goodness, Dice and Jaccard measures. The isometrics of goodness are lines whose the slope is -1. The isometrics of Jaccard and Dice are curves whose the inflexion point lies on the line $precision = recall$. The three measures give the same importance to the precision and recall but they can be sensitive to the disproportion between *precision* and *recall*. We see that Goodness does not take into account this disproportion whereas Dice and Jaccard are very sensitive, they promote *balanced* biclusters, i.e. biclusters where *precision* and *recall* are equal. Note that $precision = recall$ implies equality of Dice and Goodness. The more we move away from the line $precision = recall$, the more we are penalized by Dice. On the other hand, the isometrics of Dice and Jaccard are similar; the figure 2 illustrates that these two measures are compatible since if we draw their isometrics on the same graphics, the lines do not cross. There is a difference in the

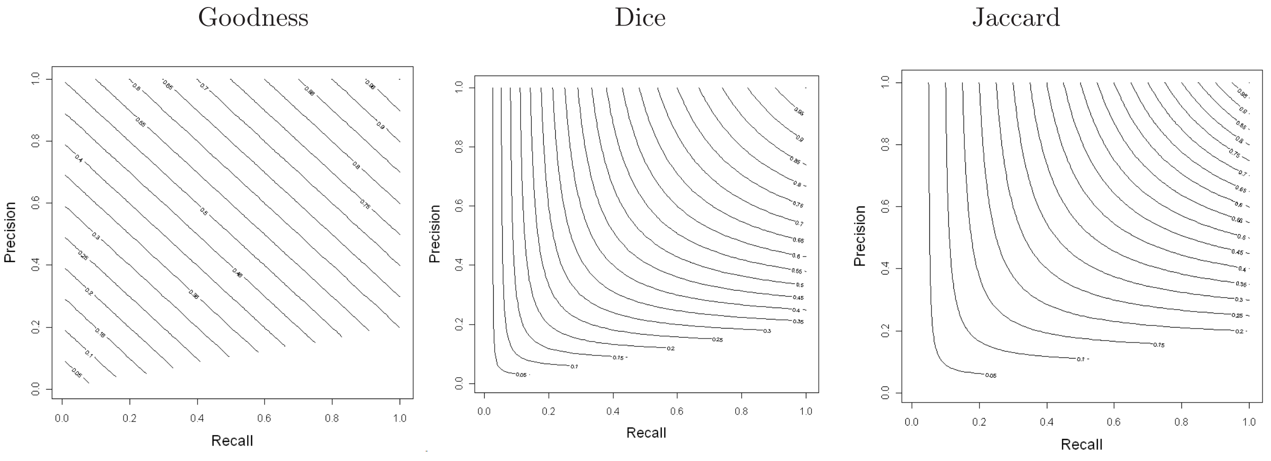

Figure 2. Isometrics of Goodness, Dice and Jaccard measure in the Precision-Recall space.

distribution of their values. The isometrics of Jaccard are more concentrated to the point (1,1) meaning that the range of values used to measure the bicluster performances is larger for good biclusters than with the Dice measure. Whereas for bad and medium biclusters the range of Jaccard is smaller than the range of Dice. This analysis leads to the following recommendations: if we do not care about the disproportion of precision and recall, we should use the Goodness measure. If we want "equilibrate" biclusters, we should use Dice or Jaccard. We will prefer the Jaccard measure in easy biclustering problems, where the estimated biclusters tend to obtain good performances. However, the Dice measure could be used in harder problems, where the estimated biclusters are not good enough.

3.2. Precision-Recall Tradeoff

All performance measures given in the previous section consider that the precision and recall have the same importance, but in practice, this is not always the case. For example, in microarray data analysis, biclustering is used to select subset of genes presenting some potentially interesting patterns. Then the elements contained in the bicluster are analyzed manually by a biologist in comparing the corresponding genes with the bibliography and making some biological experiments. Biclustering is therefore used as a preprocessing method in order to reduce the size of the data. In this context, recall is much more important than precision. A measure giving the same importance to recall and precision is therefore not suitable for this problem. A reliable performance measure should use a tradeoff between precision and recall adapted to the context. We present a variant of Goodness in introducing a parameter R that controls the tradeoff precision-recall:

$$M_{good} = \frac{1}{R+1}(R.pre + rec), \tag{9}$$

where R is the ratio of importance of precision compared to recall, for example $R = 2$ means that precision is twice more important than recall. When $R = 1$ precision and recall have the same importance and we obtain the same definition as in formulas (6). The denominator $(R + 1)$ is a normalization term such that Goodness remains in $[0, 1]$. In the previous section we have shown that Dice is actually the F1-measure. We can use the parameter β of the F-measure to control the tradeoff precision recall:

$$M_{Fmeasure}(B, X) = (1 + \beta^2)\frac{pre.rec}{\beta^2 pre + rec}, \tag{10}$$

where β is the ratio of importance of precision compared to recall. When $\beta = 1$ precision and recall have the same importance and obtain the definition of the Dice measure (4). In the rest of this paper, we replace Dice by F-measure since Dice is just a special case.

4. Size Bias Correction

In the comparison studies, generally we have to compare biclustering algorithms that produce estimated biclusters of different sizes. A major problem of the performance measures is that they have a bias depending on the size of estimated biclusters. We have performed some experiments to point out this bias. We have considered a 100 × 100 data matrix containing a true bicluster B of size 30 × 30 or 50 × 50. We have generated 10000 random biclusters of various sizes and computed their performance in studied measures terms. For each bicluster, we computed its

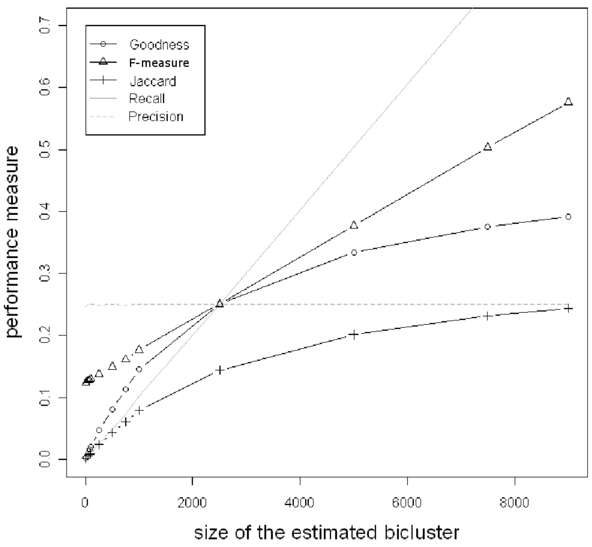

Figure 3. Performance measure for random biclustering.

precision, recall, Goodness ($R = 1$), Jaccard and F-measure ($\beta = 1$).

The figure 3 shows the average of the performance measures in function of the size of the biclusters, the size of the true bicluster is $|B| = 50 \times 50$. Full and dotted gray lines represent recall and precision. We see that the precision does not depend on the size of the estimated bicluster, its line is constant and equal to $\frac{|B|}{|D|} = \frac{2500}{10000} = 0.25$. The recall increases linearly with $|X|$, the slope of the line is $\frac{|1|}{|D|} = 10^{-4}$. The circle, cross and triangle curves represent Goodness, Jaccard and F-measure respectively. Without surprise we observe that Goodness increases linearly with $|X|$ since it is a linear combination of recall and precision. Jaccard and F-measure are increasing with $|X|$. The increase is strong for small estimated biclusters and weak for large estimated biclusters. Jaccard measure tends to the precision when $|X|$ tends to $|D|$. Since all biclusters are randomly chosen, we expected that all of them obtain the same performance, but this is not the case. This figure shows that there is a size bias for all measures (excepted precision). The large biclusters are at an advantage compared to small biclusters. The consequence of this bias is that the comparison of different biclustering algorithms producing biclusters of different sizes is not reliable and may lead to wrong conclusions.

5. Corrected Measures

We propose some modifications of the different biclustering measures in order to remove the effect of size bias. Our approach is to apply the following correction:

$$M(B,X)^{correct} = \frac{M(B,X) - \quad [M(B,X)]}{1 - \quad [M(B,X)]}$$

where $[M(B,X)]$ corresponds to the expected value of the measure M for a random bicluster of size $|X|$. The subtraction by $[M(B,X)]$ removes the effect of the size bias, the denominator adjusts the range of the measure such that 1 corresponds to the perfect bicluster and 0 to the worst performance, i.e. equivalent to a random bicluster. Note that the corrected measure is actually defined in [-1,1], but the negative values means that the biclusters are worse than random biclusters. The performance of this kind of biclusters can be considered equal to 0 and the corrected measure defined in the range $[0, 1]$. This type of correction has already been used in some works (Lee et al., 2011), but its computation was empirical. It was estimated by the average of measures obtained by generating a large set of random biclusters. This method has some drawbacks. It is very time consuming since we need to generate and compute the performance of several hundreds or thousands of biclusters. Moreover, since we do not know the variance of $M(B,X)$, the minimum number of iterations to obtain a reliable estimation of $[M(B,X)]$ is unknown, so this estimation may be inaccurate. In the following, we present an analytical formulation of this correction and show how to compute it quickly and accurately.

In the previous sections, we have already shown that the expected precision and recall for a random bicluster of size $|X|$ are $[pre] = \frac{|B|}{|D|}$ and $[rec] = \frac{|X|}{|D|}$ respectively. Hereafter we describe the expectation of Goodness, F-measure and Jaccard measures.

Property 1 The expected Goodness for a random bicluster of size $|X|$ is defined by $\frac{R|X|+|B|}{|D|(R+1)}$.

Proof:

$$\begin{aligned}
[M_{goodness}(B,X)] &= [\frac{1}{R+1}(R.pre + rec)] \\
&= \frac{1}{R+1}(R\frac{|B|}{|D|} + \frac{|X|}{|D|}) \\
&= \frac{R|B| + |X|}{|D|(R+1)}.
\end{aligned}$$

Property 2 The expected F-measure for a random bicluster of size $|X|$ is defined by $(1+\beta^2)\frac{|B||X|}{|D|(\beta^2|B|+|X|)}$.

Proof:

For a given bicluster size $|X|$, precision and recall are totally correlated and we have $pre = \frac{|B|}{|X|}rec$, then

$$
\begin{aligned}
[M_{Fmeasure}(B,X)) &= \left[(1+\beta^2)\frac{pre.rec}{\beta^2 pre + rec}\right] \\
&= \left[(1+\beta^2)\frac{\frac{|B|}{|X|}rec^2}{\beta^2\frac{|B|}{|X|}rec + rec}\right] \\
&= (1+\beta^2)\frac{|X|}{\beta^2|B|+|X|}\frac{|B|}{|X|} \quad [rec] \\
&= (1+\beta^2)\frac{|B||X|}{|D|(\beta^2|B|+|X|)}.
\end{aligned}
$$

Property 3 The expected Jaccard measure for a random bicluster of size $|X|$ can be approximated by $\frac{|B||X|}{|D||B|+|D||X|-|B||X|}$.

Proof:

$$
\begin{aligned}
[M_{jaccard}(B,X))] &= \left[\frac{pre.rec}{pre + rec - pre.rec}\right] \\
&= \left[\frac{\frac{|B|}{|X|}rec^2}{\frac{|B|}{|X|}rec + rec - \frac{|B|}{|X|}rec^2}\right] \\
&= \left[\frac{|B|rec}{|B|+|X|-|B|rec}\right].
\end{aligned}
$$

The computation of this expectation being not tractable we propose to approximate it by

$$
\frac{|B| \quad [rec]}{|B|+|X|-|B| \quad [rec]} = \frac{|B||X|}{|D||B|+|D||X|-|B||X|}.
$$

The three panels of the figure 4 show respectively the isometrics of corrected Goodness ($R = 1$), F-measure ($\beta = 1$) and Jaccard where $|D| = 10000$ and $|B| = 2000$. Around the point (1,1) the isometrics of corrected measures are close to the isometrics of uncorrected measures. The isometrics representing $M(B,X) = 0$ lies on the horizontal line of random bicluster defined on the figure 1. We see that the isometrics of corrected measures are more complex than the uncorrected measures. The visualization of the results in the precision-recall space becomes more difficult. That is a drawback of the application of these corrections to the performance measures, the interpretability of the results decreases.

6. Corrected precision-recall space

In order to combine the advantages of corrected measures and the simple interpretation and visualization of uncorrected measures, we propose to represent the results in a new space. We define the corrected precision-recall space from the corrected recall and corrected precision.

$$
\begin{aligned}
pre^{correct} &= \frac{pre-\frac{|B|}{|D|}}{1-\frac{|B|}{|D|}} = \frac{|D|pre-|B|}{|D|-|B|} \\
rec^{correct} &= \frac{rec-\frac{|X|}{|D|}}{1-\frac{|X|}{|D|}} = \frac{|D|rec-|X|}{|D|-|X|}.
\end{aligned}
\tag{11}
$$

The figure 5 depicts the isometrics of precision and recall in the precision-recall space (top) and their transformation in the corrected precision-recall space (bottom). This figure illustrates the deformation of the space when corrected precision and recall are used. In the corrected-space all points representing performances of random biclusters, i.e. points lying on the dotted lines in uncorrected-space, have been moved to the point (0,0). The points (0,0) and (1,1) represent respectively the worst and best performance. The gray area of uncorrected-space vanishes, all points of the corrected-space are possible.

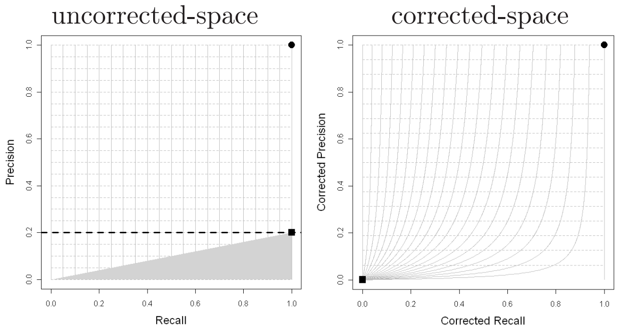

Figure 5. Comparison of the Precision-Recall space (left) and corrected Precision-Recall space (right).

From this new space, we define corrected-space measures. A corrected-space measure is computed with the uncorrected formulas of the measure but the corrected precision and recall are used instead of the uncorrected precision and recall. For example the corrected-space F-measure is computed from the formula (10) in using the corrected precision and recall defined in (11). In analysing the corrected-space Goodness, Jaccard and Dice measure, we find out some interesting properties:

Property 4 The isometrics of corrected-space Goodness, Jaccard and Dice measure in the corrected precision-recall space are exactly the same as the uncorrected Goodness, Jaccard and Dice measure in the precision-recall space (see figure 2).

Property 5 The corrected-space F-measure is equal to the corrected F-measure.

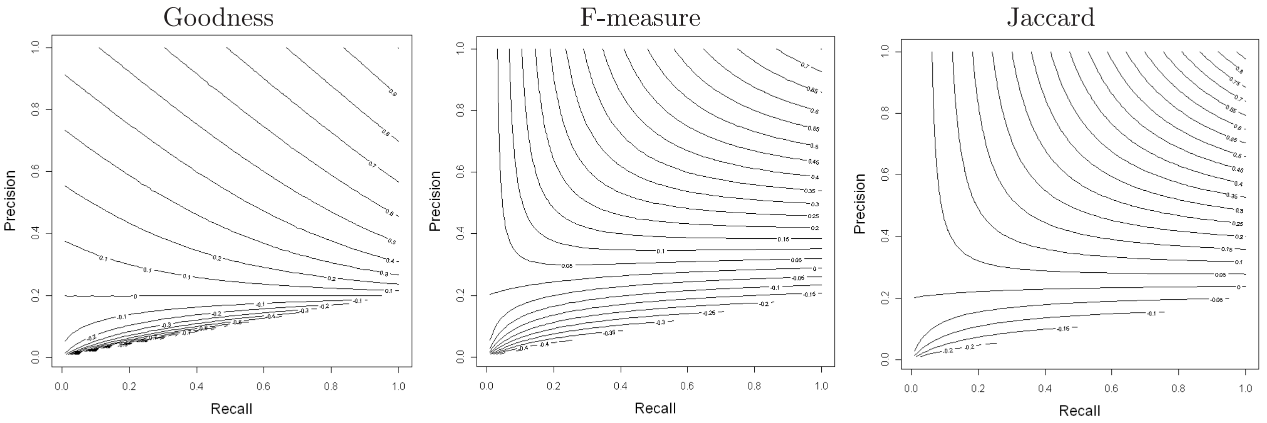

Figure 4. Isometrics of corrected Goodness, F-measure and Jaccard in the precision-recall space.

Property 6 The corrected-space Jaccard measure is compatible with the corrected Jaccard measure.

Property 7 The Jaccard and F-measure are compatible in the corrected-space

All of these propositions can be easily demonstrated in replacing the precision and recall by the corrected precision and recall in the formulas of Jaccard (7), Goodness (9) and F-measure (10). Notice that there are no propositions about the relation between corrected-space goodness and corrected goodness. We can demonstrate that these two measures are theoretically not compatible, but we will see in the next section that, in practice, corrected-space Goodness and corrected goodness are compatible in the most part of the cases. These two measures diverge only in the extreme cases where the estimated bicluster is very large.

7. Experiments

We will show the interest of our approach in order to obtain a reliable performance measure of biclusters and biclustering algorithms. These experiments have two objectives. The first one is to point out the significant difference between the uncorrected measures and the corrected, correct space measures. The second is to show that results of the corrected and corrected-space measures are compatible. This point has a high impact on the model selection and algorithm comparison.

In our experiment, a 100×100 artificial data matrix is generated, in which a 40×20 bicluster is included using the same model described in the Cheng and Church's paper (Cheng & Church, 2000). We use the popular Cheng and Church algorithm to identify an estimated bicluster in this data matrix. This algorithm depends on the parameter δ representing the maximum allowed mean square error in the estimated bicluster. The size and the quality of the estimated bicluster strongly depend on this parameter. The choice of the δ's value is therefore critical. Generally, the value that maximizes the biclustering measure is chosen. The figure 6 illustrates that using an uncorrected measure leads to a different value of δ than the corrected and corrected-space, a suboptimal bicluster is

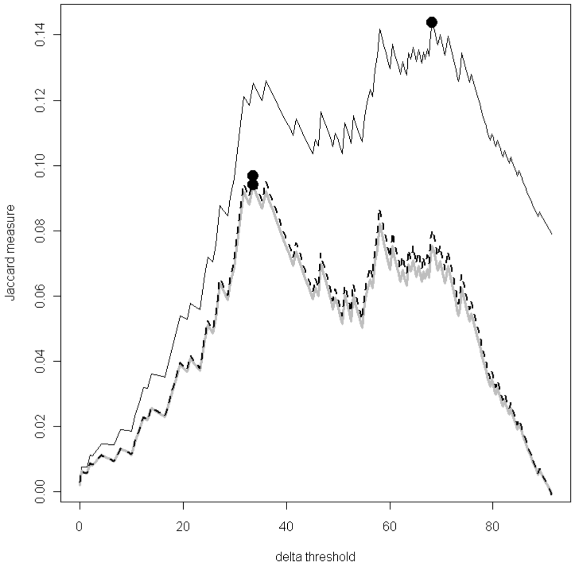

Figure 6. Jaccard measure in function of the delta threshold with the Cheng and Church algorithm. The uncorrected measure is represented by the full line, the corrected measure by the dotted lines and the corrected-space measure by the gray line. The blacks dots represent the maximal measure and the optimal threshold for each measure.

Table 1. Three comparisons of different estimated biclusters with corrected (cor.) and uncorrected (uncor.) measures according to various sizes of biclusters.

		Goodness			Jaccard			F-measure		
	size	uncor.	cor.	cor.-space	uncor.	cor.	cor.-space	uncor.	cor.	cor.-space
X_1	800	**0.70**	**0.67**	**0.67**	**0.54**	**0.52**	**0.51**	**0.70**	**0.67**	**0.67**
X_2	800	0.62	0.59	0.59	0.45	0.43	0.42	0.62	0.59	0.59
X_3	1600	0.49	0.45	0.46	0.28	**0.26**	**0.25**	0.43	**0.40**	**0.40**
X_4	266	**0.60**	**0.52**	**0.55**	**0.29**	0.24	0.23	**0.45**	0.38	0.37
X_5	3200	**0.40**	0.29	0.30	0.18	0.12	0.11	0.30	0.20	0.20
X_6	400	0.38	**0.34**	**0.34**	**0.20**	**0.18**	**0.17**	**0.33**	**0.30**	**0.30**

therefore produced. This figure shows the Jaccard measure of the estimated bicluster (y-axis) in function of the δ threshold (x-axis). The uncorrected measure is represented by the full line, the corrected measure by the dotted lines and the measure in the corrected-space by the gray line. The blacks dots represent the measure maximum and the optimal threshold for each measure. The dotted line and gray line have exactly the same shape, confirming that Jaccard and its corrected are compatible in the corrected-space. We see that the value of the optimal threshold for uncorrected Jaccard ($\delta^*_{uncor} = 67.9$) is very different than optimal threshold of corrected and corrected-space measures ($\delta^*_{cor} = \delta^*_{space} = 33.5$). The estimated bicluster returned by the uncorrected measure has a size of 3294 whereas the bicluster size from corrected measures is 475. This is an example of how the size bias can be affected the biclustering algorithm and leads to a suboptimal results. We also point out that the corrected and corrected-space measures give the same optimal δ i.e. the same estimated bicluster.

We reuse the same data matrix in the next experiment to show the impact of the different measures in the bicluster comparison. We generate different estimated biclusters of various size and quality. For each of them, the uncorrected, corrected and corrected-space versions of all measures are computed. We compare all biclusters two by two and identify the best one for each measure. Since we generated 50 estimated biclusters, we have 1225 bicluster comparisons. The table 1 gives three examples of comparisons (X_1 vs X_2, X_3 vs X_4, X_5 vs X_6).

- In the first one, all measures show that X_1 is better than X_2. But note that X_1 and X_2 have the same size (equal to the size of the true bicluster), there is no size bias, all measures give therefore the same conclusion.
- In the second comparison, X_3 is a large bicluster and X_4 a small one. The uncorrected F-measure and Jaccard give X_4 as the best bicluster whereas their corrected and corrected-space versions show that X_3 is better than X_4.
- In the third comparison, the uncorrelated Goodness does not give the same conclusion that corrected and corrected-space measures. Over all comparisons, we notice that the uncorrected measures give different conclusions than corrected and corrected-space measures in almost 20% of the comparisons. In these cases the use of an uncorrected measure leads to the wrong conclusions. The corrected and corrected-space measures give the same conclusions in almost all cases (98.8%). The differences appear only with the Goodness measure with extremely large biclusters.

8. Discussion and Conclusion

In this paper, we have presented several external measures in biclustering context. The analysis of these measures on the precision-recall space shows that the choice of a given measure implies some assumptions on the biclusters. Our analysis leads us to the following recommendations: If the precision and recall have the same importance, Goodness ($R = 1$), Dice or Jaccard can be used. In the other case, the Goodness and F-measure should be preferred, the parameters R and β control the tradeoff precision-recall. As all of these measures are affected by the size bias advantaging the large biclusters, we have proposed an efficient correction of this bias for each measure. We suggest to compute the measures in the corrected precision-recall space. Our experiments have shown that the assessment of performance must be chosen carefully, if the measure is not adapted to the context of the problem, the comparison study may be biased and leads to wrong conclusions.

References

Cheng, Y. and Church, G. M. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*, 8: 93–103, 2000. ISSN 1553-0833.

Davis, Jesse and Goadrich, Mark. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pp. 233–240, 2006.

Dhillon, Inderjit S. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '01, pp. 269–274, 2001.

Flach, Peter A. The geometry of roc space: Understanding machine learning metrics through roc isometrics. In *ICML*, pp. 194–201, 2003.

Govaert, G. and Nadif, M. Block clustering with Bernoulli mixture models: Comparison of different approaches. *Computational Statistics and Data Analysis*, 52:3233–3245, 2008.

Hanczar, B. and Nadif, M. Bagging for biclustering: Application to microarray data. In *European Conference on Machine Learning*, volume 1, pp. 490–505, 2010.

Hanczar, B. and Nadif, M. Ensemble methods for biclustering tasks. *Pattern Recognition*, 45(11):3938–3949, 2012.

Handl, Julia, Knowles, Joshua, and Kell, Douglas B. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21:3201–3212, 2005.

Lazzeroni, Laura and Owen, Art. Plaid models for gene expression data. Technical report, Stanford University, 2000.

Lee, Youngrok, Lee, Jeonghwa, and Jun, Chi-Hyuck. Stability-based validation of bicluster solutions. *Pattern Recognition*, 44:252–264, 2011.

Madeira, S. C. and Oliveira, A. L. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

Prelic, Amela, Bleuler, Stefan, Zimmermann, Philip, Wille, Anja, Buhlmann, Peter, Gruissem, Wilhelm, Hennig, Lars, Thiele, Lothar, and Zitzler, Eckart. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.