

A unifying view for performance measures in multi-class prediction

Giuseppe Jurman, Cesare Furlanello*

Fondazione Bruno Kessler, Trento, Italy

Abstract

In the last few years, many different performance measures have been introduced to overcome the weakness of the most natural metric, the Accuracy. Among them, Matthews Correlation Coefficient has recently gained popularity among researchers not only in machine learning but also in several application fields such as bioinformatics. Nonetheless, further novel functions are being proposed in literature. We show that Confusion Entropy, a recently introduced classifier performance measure for multi-class problems, has a strong (monotone) relation with the multi-class generalization of a classical metric, the Matthews Correlation Coefficient. Computational evidence in support of the claim is provided, together with an outline of the theoretical explanation.

Keywords: Matthews Correlation Coefficient, Confusion Entropy, classifier performance measure.

1. Introduction

One of the major task in machine learning is the comparison of classifiers' performance. This comparison can be carried out either by means of statistical tests (Demšar, 2006; García & Herrera, 2008) or using a performance measure as an indicator to derive similarities and differences. For binary problems, a number of meaningful metrics are available and their properties are well understood. On the other hand, the definition of performance measures in the context of multi-class classification is still an open research topic, although several functions have been proposed in the last few years: see (Sokolova & Lapalme, 2009; Ferri et al., 2009) for two comparing reviews, (Felkin, 2007) for a discussion of the differences between the use of the same classifier on a binary and a multi-class task and (Diri & Albayrak, 2008) for an alternative graphical comparison approach. As an example, one of the most important measures for binary classifier, the Area Under the Curve (AUC) (Hanley & McNeil, 1982; Bradley, 1997) associated to the Receiver Operating Characteristic curve has no automatic extension to the multi-class case. Although an agreed reasonably average-based build extension exists (presented in (Hand & Till, 2001)), several alternative formulations are being presented, either based on a multi-class ROC approximation (Everson & Fieldsend, 2006; Landgrebe & Duin, 2005, 2006, 2008)) or by viewing the ROC as a surface whose volume (Volume Under the Surface, VUS) has to be computed (by exact integration or polynomial approximation) as in (Ferri et al., 2003; Van Calster et al., 2008; Li, 2009). Other measures are more naturally defined, starting from the accuracy (ACC, *i.e.* the fraction of correctly predicted samples) and the similar Global Performance Index

(Freitas et al., 2007a,b)), to the Matthews correlation coefficient (MCC). This latter function was introduced in (Matthews, 1975) and it is also known as the ϕ -coefficient, corresponding for a 2×2 contingency table to the square root of the average χ^2 statistic $\sqrt{\chi^2/n}$. MCC has recently attracted the attention of the machine learning community (Baldi et al., 2000) as one of the best method to summarize into a single value the confusion matrix of a binary classification task. Its use as one of the preferred classifier performance measure as increased since then, and for instance it has been chosen (together with AUC) as the elective metric in the US FDA-led initiative MAQC-II aimed at reaching consensus on the best practices for development and validation of predictive models based on microarray gene expression and genotyping data for personalized medicine (The MicroArray Quality Control (MAQC) Consortium, 2010). A generalization to the multi-class case was defined in (Gorodkin, 2004), later used also for comparing network topologies (Supper et al., 2007; Stokic et al., 2009). Finally, another interesting set of measures that have a natural definition for multi-class confusion matrices consists of the functions derived from the concept of (information) entropy, first introduced by Shannon in his famous paper (Shannon, 1948). Many measure have been defined in the classification framework based on the entropy function, from simpler ones such as the confusion matrix entropy (van Son, 1994), to more complex expressions as the transmitter information (Abramson, 1963) or the relative classifier information (RCI) (Sindhwani et al., 2001). A novel multi-class measure belonging to this set has been recently introduced under the name of Confusion Entropy (CEN) by Wei and colleagues in (Wei et al., 2010a,b): in this work, the authors compare their measure to RCI and accuracy, and they prove CEN to be superior in discriminative power and precision to both alternatives in terms of two statistical indicator called degree of consistency and degree of discriminatory, defined in (Huang & Ling, 2005).

*Corresponding author

Email addresses: jurman@fbk.eu (Giuseppe Jurman), furlan@fbk.eu (Cesare Furlanello)

In the present work we investigate the similarity between Confusion Entropy and Matthews correlation coefficient. In particular, we experimentally show that the two measures are strongly correlated, and their relation is globally monotone and locally almost linear. Moreover, we provide a brief outline of the mathematical links between CEN and MCC.

2. Confusion Entropy and Matthews Correlation Coefficient

Given a classification problem on S samples $\mathcal{S} = \{s_i : 1 \leq i \leq S\}$ and N classes $\{1, \dots, N\}$, define the two functions $\text{tc}, \text{pc} : \mathcal{S} \rightarrow \{1, \dots, N\}$ indicating for each sample s its true class $\text{tc}(s)$ and its predicted class $\text{pc}(s)$, respectively. The corresponding confusion matrix is the square matrix $C \in \mathcal{M}(N \times N, \mathbb{N})$ whose ij -th entry C_{ij} is the number of elements of true class i that have been assigned to class j by the classifier:

$$C_{ij} = |\{s \in \mathcal{S} : \text{tc}(s) = i \text{ and } \text{pc}(s) = j\}|.$$

The most natural performance measure is the accuracy, defined as the ratio of the correctly classified samples over all the samples:

$$\text{ACC} = \frac{\sum_{k=1}^N C_{kk}}{S} = \frac{\sum_{k=1}^N C_{kk}}{\sum_{i,j=1}^N C_{ij}}.$$

In information theory, the entropy H associated to a random variable X is the expected value of the self-information I of X :

$$H(X) = \mathbb{E}(I(X)) = \sum_{x \in X} h_b(x) = - \sum_{x \in X} p(x) \log_b(p(x)),$$

where $p(x)$ is the probability mass function of X , with the position $h_b(x) = 0$ for $p(x) = 0$, motivated by the limit $\lim_{x \rightarrow 0} x \log(x) = 0$.

The Confusion Entropy measure CEN for a confusion matrix C is defined in (Wei et al., 2010a) as:

$$\text{CEN} = \sum_{j=1}^N P_j \sum_{\substack{k=1 \\ k \neq j}}^N h_{2(N-1)}(P_{jk}^j) + h_{2(N-1)}(P_{kj}^j), \quad (1)$$

where the misclassification probabilities P are defined as the following ratios:

$$P_{ij}^j = \frac{C_{ij}}{\sum_{k=1}^N C_{jk} + C_{kj}} \quad P_{ii}^i = 0$$

$$P_{ij}^i = \frac{C_{ij}}{\sum_{k=1}^N C_{ik} + C_{ki}} \quad P_j = \frac{\sum_{k=1}^N C_{jk} + C_{kj}}{2 \sum_{k,l=1}^N C_{kl}}.$$

This measure ranges between 0 (perfect classification) and 1 for the extreme misclassification case $C_{ij} = (1 - \delta_{ij})F$, for $F \in \mathbb{N}$ (this holds for $N > 2$, while it is not true anymore for $N = 2$, see Subsec.2.1).

Let $X, Y \in \mathcal{M}(S \times N, \mathbb{F}_2)$ be two matrices where $X_{sn} = 1$ if the sample s is predicted to of class n ($\text{pc}(s) = n$) and $X_{sn} = 0$ otherwise, and $Y_{sn} = 1$ if sample s belongs to class n ($\text{tc}(s) = n$) and 0 otherwise. Using Kronecker's delta function, the definition becomes:

$$X = (\delta_{\text{pc}(s),n})_{sn} \quad Y = (\delta_{\text{tc}(s),n})_{sn}.$$

Then the Matthews Correlation Coefficient MCC can be defined as the ratio:

$$\text{MCC} = \frac{\text{cov}(X, Y)}{\sqrt{\text{cov}(X, X) \cdot \text{cov}(Y, Y)}},$$

where $\text{cov}(\cdot, \cdot)$ is the covariance function. In terms of the confusion matrix, the above equation can be written as:

$$\text{MCC} = \frac{\sum_{k,l,m=1}^N C_{kk}C_{ml} - C_{lk}C_{km}}{\sqrt{\sum_{k=1}^N \left(\sum_{l=1}^N C_{lk} \right) \left(\sum_{\substack{f,g=1 \\ f \neq k}}^N C_{gf} \right)} \sqrt{\sum_{k=1}^N \left(\sum_{l=1}^N C_{kl} \right) \left(\sum_{\substack{f,g=1 \\ f \neq k}}^N C_{fg} \right)}} \quad (2)$$

MCC lives in the range $[-1, 1]$, where 1 is perfect classification, -1 is reached in the alternative extreme misclassification case of a confusion matrix with all zeros but in two symmetric entries $C_{i,\bar{j}}, C_{\bar{j},i}$, and 0 when the confusion matrix is all zeros but for one single column (all samples have been classified to be of a class k), or when all entries are equal $C_{ij} = K \in \mathbb{N}$. In this last case, the Confusion Entropy value is $(1 - \frac{1}{N}) \log_{2N-2} 2N$; when only a single column is not zero, the Confusion Entropy can assume many different values, depending on this column's entries. Note that both measures are invariant for scalar multiplication of the whole confusion matrix.

CEN is indeed more discriminant than MCC in some situations, for instance when $\text{MCC} = 0$ as mentioned above, or when the number of samples is relatively small and thus it more likely to have different confusion matrices with the same MCC and different CEN. This can be quantitatively assessed by using the degree of discrimination introduced in (Huang & Ling, 2005): for two measures f and g on a domain Ψ , let $P = \{(a, b) \in \Psi \times \Psi : f(a) > f(b), g(a) = g(b)\}$ and $Q = \{(a, b) \in \Psi \times \Psi : f(a) = f(b), g(a) > g(b)\}$; then the degree of discriminancy for f over g is $|P|/|Q|$. For instance, in the 3-classes case with 2, 4, 3 samples respectively, the degree of discriminancy of CEN over MCC is about 6. A similar behaviour happens for all the 12 small sample size cases on three classes listed in (Wei et al., 2010a, Tab. 6), ranging from 9 to 19 samples. In the same paper (Huang & Ling, 2005), another indicator for comparing distances is defined, the degree of consistency: for two measures f and g on a domain Ψ , let $R = \{(a, b) \in \Psi \times \Psi : f(a) > f(b), g(a) > g(b)\}$ and

$S = \{(a, b) \in \Psi \times \Psi: f(a) > f(b), g(a) < g(b)\}$; then the degree of consistency of f and g is $|R|/(|R| + |S|)$.

A quite different behaviour between the two measures can be highlighted in the following situation: consider the matrix Z_A with all entries are equal but a non-diagonal one; because of the multiplicative invariance, we can set all entries to one but for the one in the leftmost lower corner: $(Z_A)_{ij} = 1 + \delta_{(i,j),(N,1)}(A - 1)$ for $A \geq 1$ a positive integer. When A grows bigger, more and more samples are misclassified: for instance, the corresponding accuracy reads $\text{ACC}(Z_A) = N/(N^2 + A - 1)$, thus decreasing towards zero for increasing A .

The MCC measure of this confusion matrix is

$$\text{MCC}(Z_A) = -\frac{A - 1}{(N - 1)(N^2 - 2A - 2)},$$

which is a function monotonically decreasing for increasing values of A , with limit $-1/(N - 1)$ for $A \rightarrow \infty$. On the other hand, the Confusion Entropy for the same family of matrices is

$$\text{CEN}(Z_A) = \frac{1}{N^2 + A - 1} [(N - 2)(N - 1) \log_{2N-2}(2N) + (2N + A - 3) \log_{2N-2}(2N + A - 1) - A \log_{2N-2}(A)],$$

which is a decreasing function of increasing A , asymptotically moving towards zero, i.e., the minimal entropy case. Thus in this case, the behaviour of the Confusion Entropy is the opposite than the one of more classical measures such as MCC and accuracy.

Analogously for the case of (perfectly) random classification on a unbalanced problem: because of the multiplicative invariance of the measures, we can assume that the confusion matrix for this case has all entries equal to one but for the last row, whose entries are all A , for $A \geq 1$. In this case, the Confusion Entropy is

$$\text{CEN} = \frac{N - 1}{2N(N + A - 1)} [(2N + A - 3) \log_{2N-2}(2N + A - 1) - 2A \log_{2N-2} A + (A + 1) \log_{2N-2}(N + NA + A - 1)],$$

which is a decreasing function for growing A whose limit for $A \rightarrow \infty$ is $\frac{N-1}{2N} \log_{2N-2} N + 1$ (as a function of N , this limit is an increasing function asymptotically growing towards $1/2$).

One of the main features of the MCC measure is the fact that $\text{MCC}=0$ identifies all those case where random classification (i.e., no learning) happens: this is lost in the case of CEN, due to its greater discriminant power - there is no unique value associated to the wide spectrum of random classification.

Consider now the confusion matrix B of dimension N where $B_{ji} = F + (T - F)\delta_{ij}$, i.e. all entries have value F but in the diagonal whose values are all T , for T, F two integers. In this case,

$$\text{MCC} = \frac{T^2 + (N - 2)TF - (N - 1)F^2}{[T + (N - 1)F]^2}$$

$$\text{CEN} = \frac{(N - 1)F}{T + (N - 1)F} \log_{2N-2} \frac{2[T + (N - 1)F]}{F},$$

and thus

$$\text{CEN} = (1 - \text{MCC}) \left(1 + \log_{2N-2} \frac{T + (N - 1)F}{(N - 1)F} \right) \left(1 - \frac{1}{N} \right).$$

This identity can be relaxed to the following generalization, which is a slight underestimate of the true CEN value:

$$\begin{aligned} \text{CEN} &\simeq \frac{1}{k} \cdot (1 - \text{MCC}) \left(1 + \log_{2N-2} \frac{\sum_{i,j=1}^N C_{ij}}{N} \right) \left(1 - \frac{1}{N} \right) \\ &\simeq \frac{1}{k} \cdot (1 - \text{MCC}) (1 - \log_{2N-2}(1 - \text{ACC})) \left(1 - \frac{1}{N} \right) \end{aligned} \quad (3)$$

where both sides are zero when $\text{MCC} = \text{ACC} = 1$, and $k = 1.012 \cdot \left(1 + \frac{0.18924}{\log(N)} - \frac{0.06694}{\log^2(N)} \right)$. For simplicity sake, we call the right member of Eq. 3 transformed MMC, tMCC for short.

To show that the relation in Eq. 3 is valid in a wide range of situations, an experiment has been performed, whose result is graphically reported in Fig. 1. In details, 200.000 confusion matrices in dimensions ranging from 3 to 30 have been generated with the following setup: the number correctly classified elements (i.e., the diagonal elements) for each class has been (uniformly) randomly chosen between 1 and 1000, while each non-diagonal entry has been chosen as a random integer between 1 and $\lfloor 1000\rho_i \rfloor$, where the ratio ρ_i for the i -th matrix M_i was extracted from the uniform distribution in the range $[0.01, 1]$. The correlation between tMCC and k -CEN is 0.9941477 and the degree of consistency is $1 - 10^{-7}$ (the degree of discriminancy is undefined since no ties occurred). In particular, the average ratio between tMMC and k -CEN is 1.000508, with 95% bootstrap Student confidence interval (1.000328, 1.000711).

2.1. The binary case

In the binary case of two classes positive (P) and negative (N), the confusion matrix becomes $\begin{pmatrix} TP & FN \\ FP & TN \end{pmatrix}$, where T and F stands for true and false respectively.

In this setup, the Matthews correlation coefficient has the following shape:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

Similarly, the Confusion Entropy can be written as:

$$\begin{aligned} \text{CEN} &= \frac{(FN + FP) \log_2((TP + TN + FP + FN)^2 - (TP - TN)^2)}{2(TP + TN + FP + FN)} \\ &\quad - \frac{FN \log_2 FN + FP \log_2 FP}{TP + TN + FP + FN}. \end{aligned}$$

Note that in the case $TP = TN = T$ and $FP = FN = F$, the Confusion Entropy reads

$$\text{CEN} = \frac{F}{T + F} \log_2 \frac{2(T + F)}{F},$$

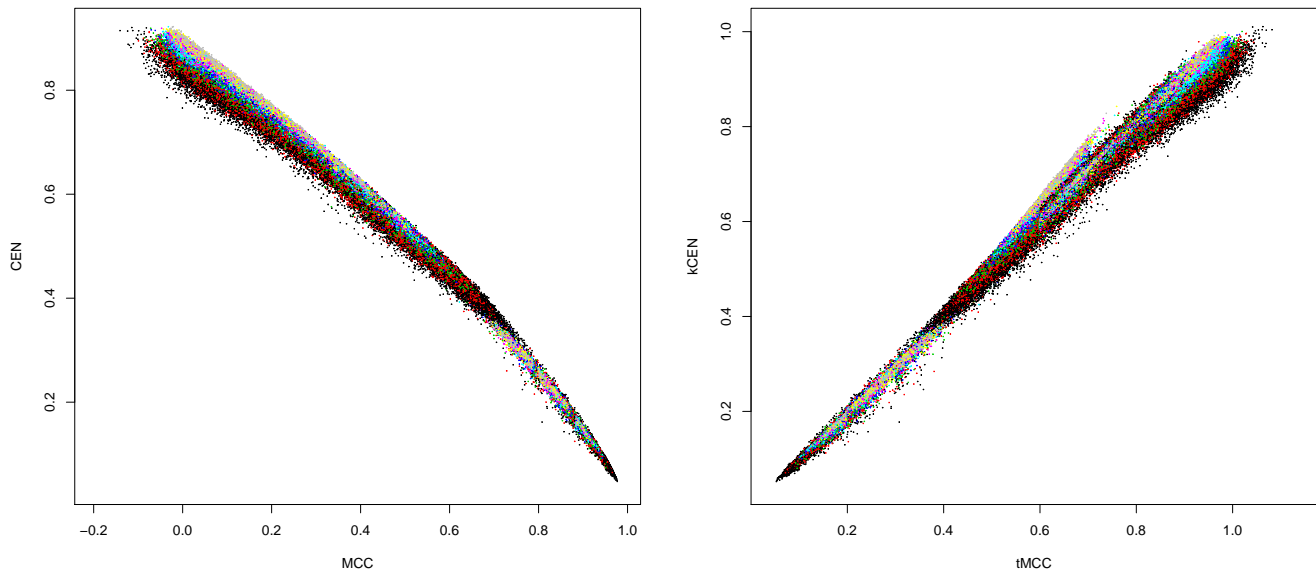


Figure 1: Plot of CEN versus MCC (left) and k -CEN versus tMCC (right) for 200.000 random confusion matrices. Each dot represents a confusion matrix, and the color indicates the matrix dimension.

which is bigger than 1 when the ratio T/F is smaller than 1. This means that all the confusion matrices $(\begin{smallmatrix} T & F \\ F & T \end{smallmatrix})$ with $0 < T < F$ have a confusion entropy larger than 1, attained for the totally misclassified case $T = 0$. Such behaviour makes CEN unusable as a classifier performance measure in the binary case.

3. Conclusions

Accuracy, Matthews Correlation Coefficient and Confusion Entropy are three crucial performance measures for evaluating the outcome of a classification task, both on binary and multi-class problems (the fourth one is Area Under the Curve, whenever a ROC curve can be drawn). Although they show a mutual consistent behaviour, each of them is better tailored to deal with different situations.

Accuracy is by far the simplest one, and its role is to convey a first rough estimate of the classifier goodness. Its use is widespread among the scientific literature, but it suffers from several caveats, the most relevant being the inability to cope with unbalanced classes and thus the impossibility of distinguish among different kinds of misclassifications.

Confusion Entropy, on the other hand, is probably the finest measure and it shows an extremely high level of discriminancy even between very similar confusion matrices. However, this feature is not always welcomed, because it makes the interpretation of its value quite harder, especially when considering situations that are naturally very similar (e.g, all the cases with $MCC=0$). Moreover, CEN may show erratic behaviour in the binary case.

In this spirit, the Matthews Correlation Coefficient is a good compromise between reaching a reasonable discriminancy degree among different cases, and the need for the practitioner of

a easily interpretable value expressing the type of misclassification associated to the chosen classifier on the given task. We showed here that there is a strong linear relation between CEN and a logarithmic function of MCC regardless of the dimension of the considered problem. Furthermore, MCC behaviour is totally consistent also for the binary case.

This given, we can suggest MCC as the best off-the-shelf evaluating tool for general purpose tasks, while more subtle measures such as CEN should be reserved for specific topic where more refined discrimination is crucial.

References

- Abramson, N. (1963). *Information theory and coding*. McGraw-Hill.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., & Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16, 412–424.
- Bradley, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30, 1145–1159.
- Demšar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, 7, 1–30.
- Diri, B., & Albayrak, S. (2008). Visualization and analysis of classifiers performance in multi-class medical data. *Expert Systems with Applications*, 34, 628–634.
- Everson, R., & Fieldsend, J. (2006). Multi-class ROC analysis from a multi-objective optimisation perspective. *Pattern Recognition Letters*, 27, 918–927.
- Felkin, M. (2007). Comparing Classification Results between N-ary and Binary Problems. In *Studies in Computational Intelligence* (pp. 277–301). Springer-Verlag volume 43.
- Ferri, C., Hernández-Orallo, J., & Modroui, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30, 27–38.
- Ferri, C., Hernández-Orallo, J., & Salido, M. (2003). Volume under the ROC surface for multi-class problems. In *In Proc. of 14th European Conference on Machine Learning* (pp. 108–120). Springer-Verlag.
- Freitas, C., De Carvalho, J., Oliveira Jr., J., Aires, S., & Sabourin, R. (2007a). Confusion matrix disagreement for multiple classifiers. In L. Rueda,

- D. Mery, & J. Kittler (Eds.), *Proceedings of 12th Iberoamerican Congress on Pattern Recognition, CIARP 2007, LNCS 4756* (pp. 387–396). Springer-Verlag.
- Freitas, C., De Carvalho, J., Oliveira Jr., J., Aires, S., & Sabourin, R. (2007b). Distance-based Disagreement Classifiers Combination. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2007* (pp. 2729–2733). IEEE.
- García, S., & Herrera, F. (2008). An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *Journal of Machine Learning Research*, 9, 2677–2694.
- Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28, 367–374.
- Hand, D., & Till, R. (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 45, 171–186.
- Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29–36.
- Huang, J., & Ling, C. (2005). Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17, 299–310.
- Landgrebe, T., & Duin, R. (2005). On Neyman-Pearson optimisation for multi-class classifiers. In *Proc. 16th Annual Symposium of the Pattern Recognition Assoc. of South Africa*. PRASA.
- Landgrebe, T., & Duin, R. (2006). A simplified extension of the Area under the ROC to the multiclass domain. In *Proc. 17th Annual Symposium of the Pattern Recognition Assoc. of South Africa* (pp. 241–245). PRASA.
- Landgrebe, T., & Duin, R. (2008). Efficient multiclass ROC approximation by decomposition via confusion matrix perturbation analysis. *IEEE Transactions Pattern Analysis Machine Intelligence*, 30, 810–822.
- Li, Y. (2009). *A generalization of AUC to an ordered multi-class diagnosis and application to longitudinal data analysis on intellectual outcome in pediatric brain-tumor patients*. Ph.D. thesis College of Arts and Sciences, Georgia State University.
- Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta - Protein Structure*, 405, 442–451.
- Shannon, C. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27, 379–423, 623–656.
- Sindhvani, V., Bhattacharge, P., & Rakshit, S. (2001). Information theoretic feature crediting in multiclass Support Vector Machines. In R. Grossman, & V. Kumar (Eds.), *Proc. First SIAM International Conference on Data Mining, ICDM01* (pp. 1–18). SIAM.
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45, 427–437.
- van Son, R. (1994). *A method to quantify the error distribution in confusion matrices*. Technical Report IFA Proceedings 18 Institute of Phonetic Sciences, University of Amsterdam.
- Stokic, D., Hanel, R., & Thurner, S. (2009). A fast and efficient gene-network reconstruction method from multiple over-expression experiments. *BMC Bioinformatics*, 10, 253.
- Supper, J., Spieth, C., & Zell, A. (2007). Reconstructing Linear Gene Regulatory Networks. In E. Marchiori, J. Moore, & J. Rajapakse (Eds.), *Proceedings of the 5th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBIO2007, LNCS 4447* (pp. 270–279). Springer-Verlag.
- The MicroArray Quality Control (MAQC) Consortium (2010). The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28, 827–838.
- Van Calster, B., Van Belle, V., Condous, G., Bourne, T., Timmerman, D., & Van Huffel, S. (2008). Multi-class AUC metrics and weighted alternatives. In *Proc. 2008 International Joint Conference on Neural Networks, IJCNN08* (pp. 1390–1396). IEEE.
- Wei, J.-M., Yuan, X.-J., Hu, Q.-H., & Wang, S.-Q. (2010a). A novel measure for evaluating classifiers. *Expert Systems with Applications*, 37, 3799–3809.
- Wei, J.-M., Yuan, X.-J., Yang, T., & Wang, S.-Q. (2010b). Evaluating Classifiers by Confusion Entropy. *Information Processing & Management, Submitted*.