Comparing Correlation Coefficients as Dissimilarity Measures for Cancer Classification in Gene Expression Data

Pablo A. Jaskowiak and Ricardo J. G. B. Campello

Department of Computer Sciences
University of São Paulo at São Carlos
São Carlos, Brazil
{pablo,campello}@icmc.usp.br

Abstract. An important analysis performed in gene expression data is sample classification, e.g., the classification of different types or subtypes of cancer. Different classifiers have been employed for this challenging task, among which the k-Nearest Neighbors (kNN) classifier stands out for being at the same time very simple and highly flexible in terms of discriminatory power. Although the choice of a dissimilarity measure is essential to kNN, little effort has been undertaken to evaluate how this choice affects its performance in cancer classification. To this extent, we compare seven correlation coefficients for cancer classification using kNN. Our comparison suggests that a recently introduced correlation may perform better than commonly used measures. We also show that correlation coefficients rarely considered can provide competitive results when compared to widely used dissimilarity measures.

Keywords: Correlation Coefficients, kNN, Cancer Classification

1 Introduction

Microarray technology enables expression level measurement for thousands of genes in a parallel fashion. The genomic picture obtained with the technology can help to discriminate among different classes of samples, which are usually associated with distinct types of cancer. Sample classification is not only essential to successful cancer diagnosis, but also to help choosing the best treatment for different patients, minimizing collateral effects of the treatment [19]. A wide range of classifiers have been applied to this task [2, 8, 12]. Among these, k-Nearest Neighbors (kNN) [5] is of main interest for our work, once it has shown quite good results in cancer classification problems [2, 8, 15]. In addition, kNN is fairly simple and straightforward to implement, which makes it very appealing.

In brief, kNN has two parameters that must be set or adjusted, which can, in turn, directly affect the classification outcomes. The first one is the number of neighbors (k), while the second one is the dissimilarity measure that induces the neighboring relationships. Parameter k and the effect of its choice in cancer

classification have been explored in some works [12, 13]. In what concerns the dissimilarity measure, on the other hand, Euclidean distance and Pearson correlation have been widely adopted as rules of thumb [2, 8, 12]. Adding to the fact that these measures are not the only alternatives to quantify dissimilarities in gene expression data, kNN can be quite sensitive to the dissimilarity choice [1]. In spite of that, little effort has been made to try establishing guidelines to the choice of a proper dissimilarity measure in this particular context. Parry et al. [13] evaluated three different dissimilarities with kNN. The evaluation, however, was based on a few datasets and did not consider any correlation coefficient, regardless their common use in gene expression data.

In the present work, to the best of our knowledge, we present the first comparison of correlation coefficients as dissimilarity measures for cancer classification using kNN. We evaluate seven correlation coefficients on 35 publicly available datasets, from both single and double channel microarrays [16]. Our investigation is motivated by sensitivity differences exhibited by the measures, such as, robustness to outliers. In addition, the correlations considered in our analysis take into account different characteristics of the data, as will be discussed later.

The remainder of this paper is organized as follows. In Section 2 the correlation coefficients considered for comparison are reviewed. In Section 3 the experimental setup is presented, whereas in Section 4 the results obtained are discussed. Finally, in Section 5 the main conclusions of our work are addressed.

2 Correlation Coefficients

When considering gene expression data and comparing any two objects (samples in our case), it turns out that these objects should be regarded as similar if they exhibit similarity in shape, rather than in absolute differences from their values [19]. Correlation coefficients have been widely used in this context, since they capture such a kind of similarity. In this sense, any two samples can be seen as sequences of real values $\mathbf{a}$ and $\mathbf{b}$, in the form $\mathbf{a} = (a_1, \ldots, a_n)$ and $\mathbf{b} = (b_1, \ldots, b_n)$, for which a correlation coefficient can be directly applied. Such coefficients produce values between -1 and 1. High absolute values indicate a stronger relationship between sequences, while values close to 0 indicate non-correlated sequences. Bearing the above considerations in mind, in the following we review the seven correlation coefficients considered in our comparison.

2.1 Pearson - ρ

The Pearson correlation [14] allows the identification of linear correlations between two sequences of numbers. It is described in (1), where $\bar{a}$ and $\bar{b}$ stand for the means of the sequences in hand. Although widely used in gene expression data, Pearson is sensitive to the presence of outliers and may not be robust when both sequences do not come from an approximately normal distribution [19, 10].

$$\rho(\mathbf{a}, \mathbf{b}) = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} \tag{1}$$

2.2 Jackknife - ϱ

Jackknife correlation [10] is defined in order to minimize the effect that single outliers may have in the final correlation value. This is achieved by removing one value at a time from both sequences. Jackknife is defined in (2), where $\rho^i(\mathbf{a}, \mathbf{b})$ stands for the Pearson correlation between $\mathbf{a}$ and $\mathbf{b}$ with their i^{th} values removed.

$$\varrho(\mathbf{a}, \mathbf{b}) = min\{\rho^1(\mathbf{a}, \mathbf{b}), \ldots, \rho^n(\mathbf{a}, \mathbf{b}), \rho(\mathbf{a}, \mathbf{b})\} \tag{2}$$

2.3 Goodman-Kruskal - γ

The Goodman-Kruskal correlation [9] takes into account only the ranks of sequences $\mathbf{a}$ and $\mathbf{b}$, and is defined according to the number of concordant (S_+), discordant (S_-), and neutral pairs in the sequences. A pair is said to be concordant if the same relative order applies in the two sequences ($a_i < a_j$ and $b_i < b_j$ or $a_i > a_j$ and $b_i > b_j$). Similarly, discordant pairs are those in which the inverse relative order applies ($a_i < a_j$ and $b_i > b_j$ or $a_i > a_j$ and $b_i < b_j$). All other pairs are deemed as neutrals. Goodman-Kruskal is defined by Equation (3).

$$\gamma(\mathbf{a}, \mathbf{b}) = \frac{S_+ - S_-}{S_+ + S_-} \tag{3}$$

2.4 Kendall - τ

The Kendall correlation [11] is based on the same concepts previously defined for Goodman-Kruskal. The difference between these two measures is due to the fact that Kendall, defined in (4), takes into account all the $n(n-1)/2$ pairs in its normalization term, reaching its extrema only in the absence of neutrals.

$$\tau(\mathbf{a}, \mathbf{b}) = \frac{S_+ - S_-}{n(n-1)/2} \tag{4}$$

2.5 Spearman - $\hat{\rho}$

The Spearman correlation [17] can be seen as a particular case of the Pearson correlation, provided that the values of both $\mathbf{a}$ and $\mathbf{b}$ are replaced with their ranks in the respective sequences. By doing such a replacement, the Spearman correlation can also be defined by (1). As only the ranks of the sequences are considered, Spearman is more robust to the presence of outliers than Pearson [19].

2.6 Rank-Magnitude - r

The Rank-Magnitude correlation [4] was introduced as an asymmetric measure, for cases in which one of the sequences is composed by ranks and the other by real values. The correlation is defined by (5), with $R(a_i)$ denoting the rank of the i^{th} element of sequence $\mathbf{a}$. In (5), $r^{min} = \sum_{i=1}^{n}(n+1-i)\bar{b}_i$ and $r^{max} = \sum_{i=1}^{n} i\bar{b}_i$.

Value $\overline{b}_i$ is the i^{th} element of the sequence, which is obtained by rearranging sequence $\mathbf{b}$ so that it gets sorted in ascending order.

$$\hat{r}(\mathbf{a}, \mathbf{b}) = \frac{2\sum_{i=1}^{n} R(a_i)b_i - r^{max} - r^{min}}{r^{max} - r^{min}} \tag{5}$$

As previously mentioned, Rank-Magnitude is asymmetric. To be used in cases in which both sequences are constituted of real values, the measure must be symmetrized. Any mention to the Rank-Magnitude correlation in the remainder of this paper refers to its symmetric version, given by $r(\mathbf{a}, \mathbf{b}) = (\hat{r}(\mathbf{a}, \mathbf{b}) + \hat{r}(\mathbf{b}, \mathbf{a}))/2$.

2.7 Weighted Goodman-Kruskal - $\hat{\gamma}$

The Weighted Goodman-Kruskal correlation [4] takes into account ranks and magnitudes of both sequences by considering that concordance and discordance are both a matter of degree. The complete Weighted Goodman-Kruskal formulation is presented in [4] and is omitted here due to space restrictions.

3 Experimental Setup

We compared different variants of the kNN classifier, each of which employing one of the correlation coefficients described in Section 2. All correlation coefficients were adapted as dissimilarities in the form: $Dissimilarity(\mathbf{a}, \mathbf{b}) = 1 - correlation(\mathbf{a}, \mathbf{b})$. For completeness, we also included the Euclidean distance (represented by letter 'e') in our comparison. Regarding kNN parameter k, we considered four values during our evaluation: 1 (1NN), 3 (3NN), 5 (5NN) and 7 (7NN). These values were chosen based on the work of Dudoit et al. [8], in which the authors show that for small sample cancer classification values of k smaller than 7 are usually preferred. Each kNN variant was evaluated by its generalization capability (error rates), which was estimated using Leave One Out Cross Validation (LOOCV). It is worth noticing that the choice of error estimators in the case of small sample sizes is still under debate [3, 7].

To evaluate each one of the 32 kNN variants considered we used a set of publicly available benchmark datasets proposed in [16]. Briefly, this benchmark set encompass 35 microarray datasets from cancer gene expression experiments and comprehend the two flavors in which the technology is generally available: single channel (21 datasets) and double channel (14 datasets) [19, 18]. Hereafter we refer to single channel microarrays as Affymetrix and double channel microarrays as cDNA, since the data was collected using either of these technologies [16]. Detailed information about these datasets can be obtained in [16].

Finally, to provide reassurance about the validity of our results, we used the Friedman and Nemenyi statistical tests (with a 95% confidence level), which are more appropriate when comparing multiple classifiers on multiple datasets [6].

4 Results

Once we are dealing with two different microarray technologies, i.e., cDNA and Affymetrix, we chose to analyze the results obtained for each technology independently. In Fig. 1 and 2 we present classification error boxplots for each one of the 32 kNN variants in cDNA and Affymetrix datasets respectively.

Considering cDNA datasets (Fig. 1), the largest variabilities were found with Euclidean distance (e), which showed the worse results among all compared measures. It is interesting to note that the use of Jackknife (ϱ) provided some decrease in variability when compared to Pearson (ρ) (except for 5NN). Regarding the rank-based correlation coefficients, Goodman-Kruskal (γ), Kendall (τ), and Spearman ($\hat{\rho}$) showed comparable results, regardless of the value of k used. These results are quite interesting and show that rank-based measures rarely used in gene expression data, such as, Goodman-Kruskal (γ) and Kendall (τ), can provide competitive results when compared to the more commonly used Spearman ($\hat{\rho}$). Considering Rank-Magnitude (r), comparable results were observed in comparison to rank-based correlations.

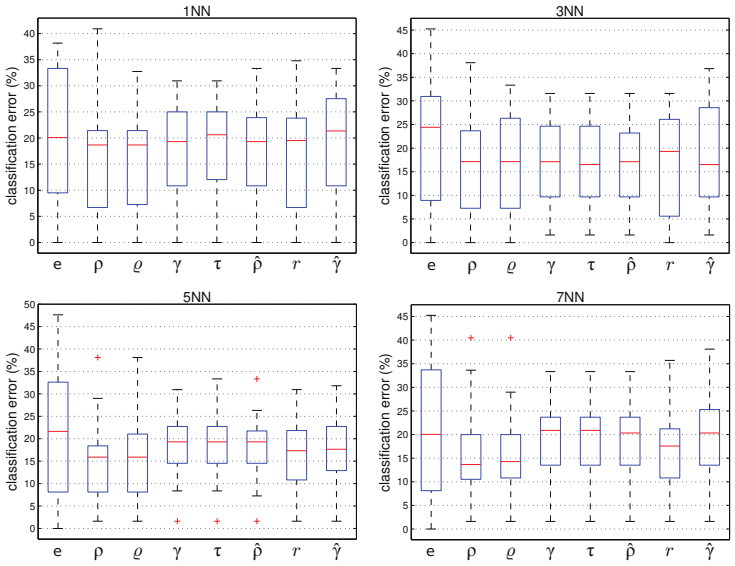

Fig. 1. cDNA datasets - boxplots showing classification errors obtained when comparing kNN with different correlation coefficients (Euclidean distance also included).

The results obtained for Affymetrix datasets are shown in Fig. 2. All measures produced outliers but the Euclidean distance (e). These outliers can, to some extent, be justified by the large number of classes present in some datasets. In this sense, the observed outliers approximate the errors that would be found

with a majority voting classifier, i.e., a classifier that assigns each unlabeled sample to the majority class observed in the training data. Regardless of the value of k, Weighted Goodman-Kruskal ($\hat{\gamma}$) displayed the largest variability among the compared dissimilarity measures. Considering the rank-based correlations, Goodman-Kruskal (γ), Kendall (τ), and Spearman ($\hat{\rho}$) showed comparable results. It is interesting to note that the commonly used Euclidean distance (e), Pearson (ρ) and Jackknife (ϱ) performed slightly worse than rank-based correlations. Rank-Magnitude correlation (r) displayed the lowest variabilities among the compared measures considering 5NN and 7NN. This correlation also performed well for 1NN and 3NN, showing better results when compared to the commonly employed Pearson (ρ) and Euclidean distance (e).

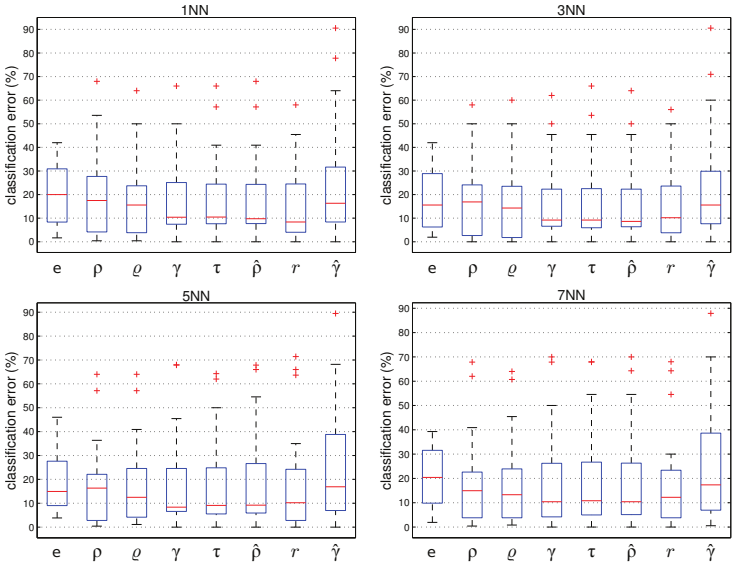

Fig. 2. Affymetrix datasets - boxplots showing classification errors obtained when comparing kNN with different correlation coefficients (Euclidean distance also included).

Fig. 1 and Fig. 2 provide an overview but hide some interesting results. In order to elucidate such cases, we also present results for each individual dataset considering only 1NN classifier[1], as shown in Table 1. In *west-2001* and *ramaswamy-2001* datasets, for which the commonly used Euclidean distance (e) and Pearson (ρ) led to larger error rates, the use of rank-based measures and Rank-Magnitude (r) can decrease their errors in almost 25%. Considering *bitner-2000*, the use of Pearson (ρ) and Jackknife (ϱ) led to a difference of almost 20% in error when compared to the Euclidean distance (e). It is important to note

[1] Results concerning the other values of k are omitted due to space restrictions.

that, in some datasets, e.g., *nutt-2003-v1, v2* and *v3*, the Euclidean distance (e) can provide better results when compared to the other measures.

Table 1. 1NN classification errors (%) for all evaluated dissimilarity measures. Shades of gray indicate relative performance of the dissimilarities (columns) for each dataset (rows). The lighter the cell, the lower the error for the respective dataset.

		e	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$
cDNA	alizadeh-2000-v1	19.1	21.4	21.4	23.8	23.8	23.8	23.8	23.8
	alizadeh-2000-v2	0.0	1.6	1.6	1.6	1.6	1.6	1.6	1.6
	alizadeh-2000-v3	17.7	19.4	19.4	14.5	14.5	16.1	19.4	14.5
	bittner-2000	34.2	13.2	13.2	18.4	21.1	18.4	15.8	29.0
	bredel-2005	12.0	18.0	18.0	18.0	18.0	18.0	16.0	20.0
	chen-2002	9.5	6.7	7.3	9.5	9.5	8.9	6.7	9.5
	garber-2001	24.2	19.7	19.7	21.2	21.2	21.2	19.7	22.7
	khan-2001	0.0	0.0	0.0	10.8	12.1	10.8	1.2	10.8
	lapointe-2004-v1	37.7	37.7	31.9	26.1	26.1	27.5	34.8	27.5
	lapointe-2004-v2	38.2	40.9	32.7	26.4	25.5	26.4	26.4	30.0
	liang-2005	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	risinger-2003	33.3	28.6	26.2	31.0	31.0	33.3	28.6	33.3
	tomlins-2006	21.2	16.4	16.4	20.2	20.2	20.2	20.2	18.3
	tomlins-2006-v2	23.9	19.6	20.7	25.0	25.0	23.9	21.7	23.9
Affymetrix	armstrong-2002-v1	8.3	2.8	1.4	1.4	1.4	1.4	0.0	6.9
	armstrong-2002-v2	11.1	4.2	5.6	4.2	5.6	5.6	4.2	9.7
	bhattacharjee-2001	14.3	13.3	13.8	7.9	6.4	6.4	8.4	8.9
	chowdary-2006	1.9	2.9	2.9	3.9	4.8	3.9	2.9	3.9
	dyrskjot-2003	20.0	17.5	17.5	20.0	22.5	20.0	20.0	15.0
	golub-1999-v1	6.9	4.2	2.8	8.3	8.3	8.3	2.8	5.6
	golub-1999-v2	8.3	5.6	4.2	8.3	8.3	8.3	4.2	6.9
	gordon-2002	1.7	2.2	2.2	0.0	0.0	0.0	0.0	0.0
	laiho-2007	13.5	24.3	24.3	10.8	10.8	10.8	18.9	10.8
	nutt-2003-v1	42.0	68.0	64.0	66.0	66.0	68.0	58.0	64.0
	nutt-2003-v2	32.1	53.6	46.4	50.0	57.1	57.1	39.3	57.1
	nutt-2003-v3	31.8	45.5	50.0	45.5	40.9	40.9	45.5	40.9
	pomeroy-2002-v1	32.4	38.2	35.3	17.7	17.7	17.7	32.4	17.7
	pomeroy-2002-v2	23.8	28.6	21.4	26.2	23.8	23.8	23.8	28.6
	ramaswamy-2001	34.2	27.4	22.6	24.7	26.3	25.8	18.4	90.5
	shipp-2002-v1	22.1	22.1	15.6	10.4	10.4	9.1	6.5	22.1
	singh-2002	23.5	26.5	23.5	22.6	23.5	23.5	19.6	21.6
	su-2001	15.5	11.5	10.3	10.3	8.1	9.8	5.2	11.5
	west-2001	30.6	16.3	14.3	6.1	8.2	8.2	8.2	16.3
	yeoh-2002-v1	2.0	0.4	0.4	9.3	10.5	8.9	3.6	17.3
	yeoh-2002-v2	25.8	23.8	21.4	40.7	38.3	38.3	26.6	77.8
		e	ρ	ϱ	γ	τ	$\hat{\rho}$	r	$\hat{\gamma}$

Finally, the statistical tests suggest that for Affymetrix datasets, Rank-Magnitude was statistically superior to Pearson, when considering 1NN. Still concerning Affymetrix datasets, both Rank-Magnitude and Jackknife were statistically superior to Euclidean distance, when considering 7NN. Regarding the results for cDNA datasets, no statistically significant differences were found.

5 Conclusions

We presented a comparison of seven correlation coefficients for cancer classification with kNN. Although no correlation performed best in all datasets, some interesting results were found. First of all, we showed that there are competitive alternatives to Euclidean distance and Pearson in both cDNA and Affymetrix datasets. Considering only Affymetrix data, the recently proposed Rank-Magnitude is, in some cases, statistically superior to Euclidean distance and Pearson, common choices in gene expression analysis. Regarding rank-based correlations, our results suggest that Goodman-Kruskal and Kendall are possible alternatives to the more commonly used Spearman. As in individual datasets

large differences were found with different correlations, in real applications an exploratory analysis considering different measures is seemingly the best choice.

As future work, it would be interesting to investigate possible relations between characteristics of the datasets and the results produced by the correlations coefficients in particular cases, e.g., *west-2001*, where greater differences among results from different coefficients were observed.

Acknowledgments. The authors would like to acknowledge the Brazilian research agencies CAPES, CNPq and FAPESP for financial support to this work.

References

1. Aha, D.W., Kibler, D., Albert, M.K.: Instance-based learning algorithms. Mach. Learn. 6(1), 37–66 (1991)
2. Ben-Dor, A., et al.: Tissue classification with gene expression profiles. J. Comput. Biol. 7(3-4), 559–583 (2000)
3. Boulesteix, A.L., Strobl, C., Augustin, T., Daumer, M.: Evaluating microarray-based classifiers: An overview. Cancer Informatics 6, 77–97 (2008)
4. Campello, R.J.G.B., Hruschka, E.R.: On comparing two sequences of numbers and its applications to clustering analysis. Inform. Sciences 179(8), 1025–1039 (2009)
5. Cover, T., Hart, P.: Nearest neighbor pattern classification. IEEE T. Inform. Theory 13(1), 21 – 27 (Jan 1967)
6. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006)
7. Dougherty, E.R., Sima, C., Hua, J., Hanczar, B., Braga-Neto, U.M.: Performance of error estimators for classification. Curr. Bioinf. 5, 53–67 (2010)
8. Dudoit, S., et al.: Comparison of discrimination methods for the classification of tumors using gene expression data. J. Amer. Stat. Assoc. 97(457), 77–87 (2002)
9. Goodman, L., Kruskal, W.: Measures of association for cross-classifications. J. Amer. Stat. Assoc. 49, 732764 (1954)
10. Heyer, L.J., Kruglyak, S., Yooseph, S.: Exploring expression data: Identification and analysis of coexpressed genes. Genome Res. 9(11), 1106–1115 (1999)
11. Kendall, M.G.: Rank correlation methods. Griffin, London, 4 edn. (1970)
12. Lu, J., et al.: MicroRNA expression profiles classify human cancers. Nature 435(7043), 834–838 (2005)
13. Parry, R.M., et al.: k-nearest neighbor models for microarray gene expression analysis and clinical outcome prediction. Pharmac. J. 10(4), 292–309 (2010)
14. Pearson, K.: Contributions to the mathematical theory of evolution. iii. regression, heredity, and panmixia. P. Roy. Soc. Lond. A Mat. 59, 69–71 (1895)
15. Singh, D., et al.: Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1(2), 203 – 209 (2002)
16. de Souto, M., Costa, I., de Araujo, D., Ludermir, T., Schliep, A.: Clustering cancer gene expression data: a comparative study. BMC Bioinformatics 9(1), 497 (2008)
17. Spearman, C.: The proof and measurement of association between two things. Am. J. Psychol. 100(3/4), 441–471 (1904)
18. Tarca, A.L., Romero, R., Draghici, S.: Analysis of microarray experiments of gene expression profiling. Am. J. Obstet. Gynecl. 195(2), 373 – 388 (2006)
19. Zhang, A.: Advanced analysis of gene expression microarray data. World Scientific Publishing Company, 1 edn. (2006)