ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Validation of Unsupervised Clustering Methods for Leaf Phenotype Screening

Andreas Backhaus[1], Asuka Kuwabara[2], Andrew Fleming[2], Udo Seiffert[1]

1- Biosystems Engineering, Fraunhofer Institute for Factory Operation and Automation (IFF), Magdeburg - Germany

2- Department of Animal and Plant Science
University of Sheffield, Sheffield - United Kingdom

Abstract. The assessment of visible differences in leaf shape between plant species or mutants (phenotyping) plays a significant role in plant research. This paper investigates the application of unsupervised data clustering techniques for phenotype screening to find hidden common shape categories. A set of two wildtypes and seven mutations of Arabidopsis acted as a test case. K-Means, NG, GNG, SOM and ART2a were evaluated by classical validity indices and one index derived from the task at hand. K-Means showed the best results and a low agreement between classical validity measures and task constraints was found.

1 Introduction

In plants, leaves have common functions to supply energy and oxygen by means of photosynthesis; however, leaves exist in many different forms, presumably reflecting different evolutionary strategies to cope with different environments. It has been suggested that differences in margin form influence the ability of a leaf to withstand environmental stress and leaf shape has been used as a proxy for temperature estimations in paleobotany [1].

It has been well known that gene networks as well as environmental cues control leaf shapes in a highly orchestrated manner. In order to clarify the role of genes involved in leaf development, precise description, quantification and categorization of leaf phenotypes are vital; for example, if two genes are working in the same signaling pathways, or if two genes are allelic, final leaf shape of each mutant may be almost identical. Similarly, if two genes are working in different signaling pathways, their double mutant phenotype may be more severe [2, 3, 4].

For this paper, nine genotypes of *Arabidopsis thaliana* are considered. The task constraints are twofold. *(i) Shape Partitioning:* An unknown number of shape groups which partition due to similarity in leaf shape have to be found. A solution to this problem is unsupervised clustering. We tested a number of well known clustering techniques which solve this constraint due to their inherent similarity measure applied to shape properties. A number of validity indices ranked clustering results. *(ii) Genotype Fragmentation:* If a genotype highly contributed to create certain leaf characteristics, then this genotype should be found in lesser number of clusters formed by the unsupervised clustering technique (ideally one genotype is found only in one cluster). Therefore it is desired to partition the data set so that a cluster contains a mutant line in its entirety while one cluster should also be able to hold more than one genotype. This constraint is manually evaluated and we suggest a validity index which directly incorporates this task constraint.

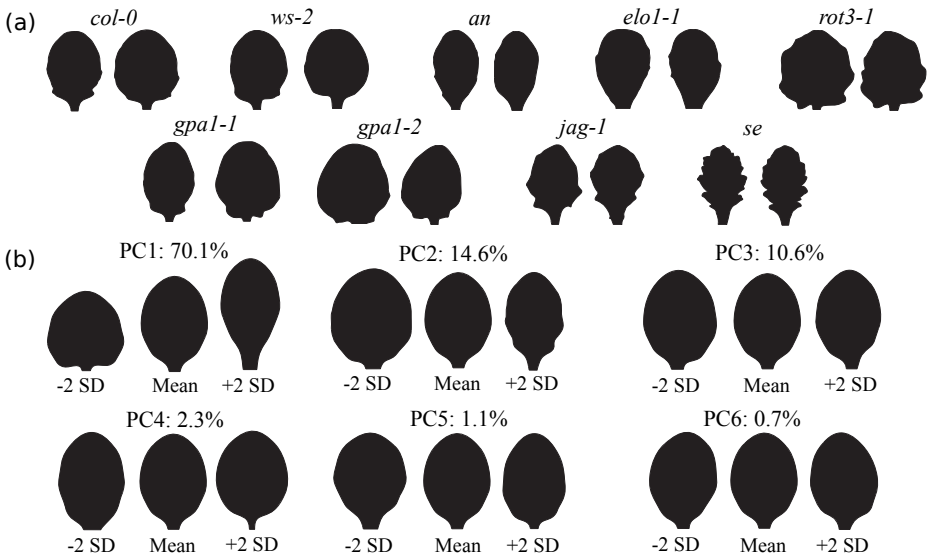

Fig. 1: (a) Two wild types (*col-0* and *ws-2*) and seven mutations of Arabidopsis have been used for the cluster analysis and validation. (b) The principal shapes which span the 6 dimensional subspace for cluster partition; PC1 and PC2: variation elongation and serration respectively; PC3: asymmetric variation leaf width, PC4: variation base width; PC5 and PC6: shape variation in lobing & width and asymmetric leaf bending respectively.

This paper shall contribute to the development of automated analysis systems of high throughput [5] that allow reliable and replicable analysis of biological structures.

2 Image Acquisition

The following wild type and mutant lines of *Arabidopsis thaliana* (L.) Heynh. were used in this study: *col-0*, *ws-2*, *angustifolia (an)*, *elongata1 (elo1-1)*, *gpa1-1* and *gpa1-2* variants, *jagged1 (jag1)*, *rotendufolia3 (rot-3)* and *serrate (se)*. Each condition held between 14 and 24 samples. Leaves were fixed at 28 days after sowing and carefully flattened while keeping the leaf margin intact. Images of flattened leaves were taken using a CCD camera mounted on a stereomicroscope at 13919 DPI and stored as 8-bit grey value bitmap. Figure 1a shows two examples per line as polygons created by the shape quantification procedure described below.

3 Shape Quantification and Feature Space

Leaves were segmented from the background by gray value thresholding and a Canny edge detector was used to create the boundary trace. A point distribution model (PDSM) was then fitted by an Active Contour [6]. Then a plane curve

ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Table 1: Best clustering results for K-Means, NG, SOM, ART2a and GNG based on various validity measures. Decision rule per index and cluster size per validation in brackets. Best results indicated in bold.

Cluster	Validity (Cluster)				
Method	HE (min)	D (max)	DB (min)	C (min)	S (max)
K-Means	**0.161** (4)	**0.136** (8)	**0.847** (6)	**0.036** (8)	**0.609** (3)
NG	0.211 (4)	0.132 (8)	0.903 (4)	0.045 (6)	0.561 (4)
SOM	0.311 (4)	0.041 (4)	1.467 (4)	0.120 (4)	0.399 (4)
GNG	0.240 (5)	0.076 (5)	1.273 (5)	0.062 (5)	0.510 (5)
ART2a	0.392 (6)	0.019 (6)	1.896 (6)	0.405 (6)	-0.220 (6)

was created through spline interpolation and 500 points in total sampled by arc length parameterization. The shape was aligned with its longest extension along the y-axis and the base was cut where the petiol has increased by 40% from its average width. Shapes were normalized in perimeter length. The normalized PDSM's (x, y) values formed a 1,000 dimensional vector per sample for further analysis.

In order to reduce dimensionality a PCA was performed on the shape data. Six principal components of largest Eigenvalues span the feature space and cover a total of 98% in data variance. Figure 1b shows the principal components at ± 2 standard deviation of coefficient value from the mean shape. Figure 2a shows the distribution of genotypes for the first two principal components.

4 Clustering Methods and Validation

Clustering is an unsupervised process without the need of predefined classes and examples that would indicate a particular partitioning within the samples. We applied a number of clustering algorithms known from the literature. They are the K-Means [7], Neural Gas (NG) [8], Self Organizing Map (SOM) [9], Growing Neural Gas (GNG) [10] and Adaptive Resonance Theory (ART2a) [11]. In K-Means, NG and GNG, each prototype or weight vector represented one cluster. Membership was decided by minimal Euclidean distance. For the SOM, a unified distance matrix based linkage clustering technique from the SOM Toolbox [12] was used to find clusters on the topological map. In ART2a, a neuron of the recognition layer represented a data cluster and membership was decided by maximum pattern correlation.

Evaluating and assessing the results of a clustering algorithm is the main subject of cluster validity. A validity criterion judges the partition into compact and well-separated clusters. The following criteria are considered in this paper: Dunn Index (D) [13], Davies-Bouldin Index (DB) [14], C-Index (C) [15] and Silhouette Index (S) [16]. High D and S indices as well as low DB and C indices indicate well-separated compact clusters.

However, these indices reflect a very common constraint on the cluster partition. For phenotype screening, as mentioned in the introduction, a constraint of low genotype fragmentation from clustering was formulated. We evaluated the fragmentation manually (see Figure 3) but also automatically via the distribution of sample 'hits' a genotype has per cluster. Let h_{ij} be the relative frequency (or 'hits') of the i-th genotype in the j-th cluster, the following index (based on

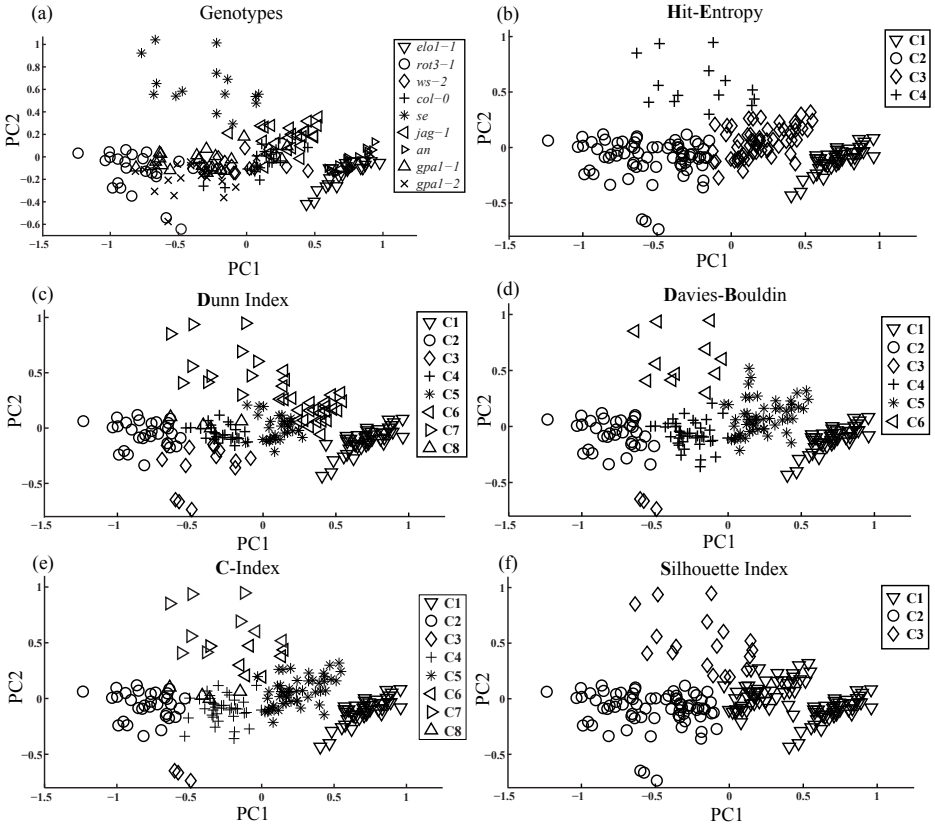

Fig. 2: (a) Shape space spanned by PC1 and PC2; (b-f) K-Means cluster partitioning selected by different validity indices.

normalized entropy)

$$HE = -\frac{1}{n_s \ln(n_c)} \sum_{i=0}^{n_s-1} \sum_{j=0}^{n_c-1} h_{ij} \ln(h_{ij}) \tag{1}$$

is minimal if genotypes are assigned to just one cluster in all their samples. Terms n_c and n_s are the number of clusters and genotype classes respectively. We shall term this index 'Hit-Entropy' (HE).

5 Results

For K-Means and NG, cluster number ranged from three to eight. For each model size, ten trials with randomized initial codebooks (K-Means, NG, GNG, SOM) were run. Cluster outputs were evaluated with the described indices. According to the indices' maximum or minimum decision rule, the best cluster configuration was found.

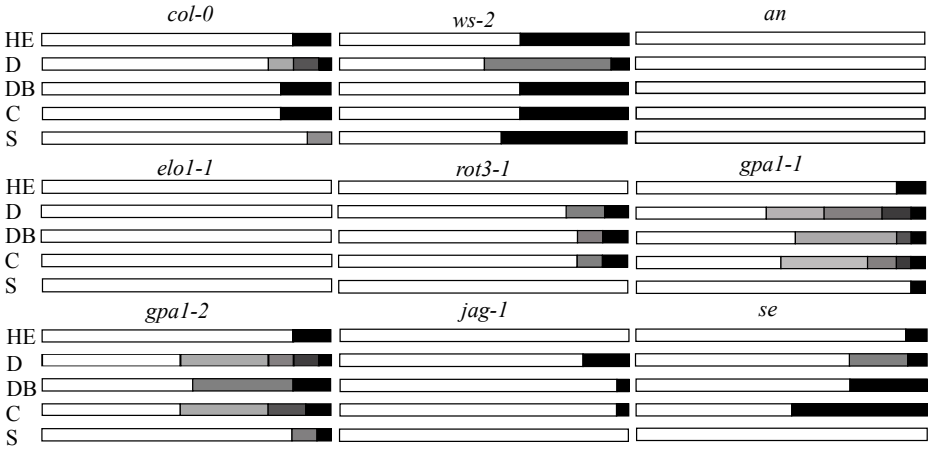

Fig. 3: The genotype class fragmentation through data clustering; Normalized hits per cluster (width segments), sorted by descending value; Number of gray shaded segments equals number of clusters a genotype is fragmented into; HE and S index showed the lowest fragmentation.

Table 1 shows the cluster performance. K-Means leads the performance table followed by NG and GNG. For the cluster model creation, the trivial cases of two and nine clusters were not considered. In preliminary tests, validity indices (besides HE) tend to choose these trivial cases in clustering results for NG and K-Means. It is interesting to note that GNG created close cluster numbers compared to 'brute force' selection based methods K-Means and NG with similar cluster partitioning (data not published here).

Shape Partitioning: Figure 2b-f shows cluster partitions selected by validity index evaluation. PC1 and PC2 basically represent shape elongation and serration respectively. The DB-index seems to select the best obvious data partitioning (five clusters). Genotypes *elo1-1* and *an* (both elongated shape) form a characteristic separate cluster as well as the *se* genotype. A group of outliers was clustered separately (DB-index). For the DB-index it is apparent that the partition is based on cluster separation. The task specific index HE chose a configuration of four clusters due to overlapping genotype classes (*gpa1-1* and *gpa1-2*).

Genotype Fragmentation: Figure 3 depicts the fragmentation of genotypes into clusters. Presented are always the winning configurations from Table 1 per index. The aim is to find a configuration with minimal fragmentation. HE, as expected, chose such a configuration followed by S index. D, DB and C indices chose fragmented configurations. This manual evaluation shows that classical validity indices not necessarily meet this task constraint.

6 Conclusion

Automated screening of biological structures generates a replicable and objective analysis compared to subjective analysis by eye sight. With the development of

high throughput imaging and growth apparatus, analysis techniques have to be developed and evaluated in parallel to motivate a still lagging advanced and systematic approach to leaf shape phenotype screening.

The evaluation of unsupervised cluster results is not a trivial task. Configurations generated from randomized initial conditions and varying number of clusters have to be evaluated due to task constraints. For the task of phenotype screening two constraints were formulated. While an index like DB was able to select a reasonable cluster partition, it failed to choose a configuration of minimal genotype fragmentation. Our task specific index HE was able to generate a compromise between shape similarity grouping and genotype fragmentation and derived an underlying cluster structure of four from the dataset.

References

[1] J.D. Krieger, R.P. Guralnick, and D.M. Smith. Generating empircially determined, continuous measures of leaf shape for paleoclimate reconstruction. *PALAIOS*, 22(2):212–219, 2007.

[2] A. Hay and M. Tsiantis. The genetic basis for differences in leaf form between *Arabidopsis thaliana* and its wild relative *Cardamine hirsuta*. *Nat Genet*, 38(8):942–947, 2006.

[3] E. Anastasiou, S. Kenz, M. Gerstung, D. MacLean, J. Timmer, C. Fleck, and M. Lenhard. Control of plant organ size by KLUH/CYP78A5-dependent intercellular signaling. *Dev Cell*, 13(6):843–856, Dec 2007.

[4] T. Blein, A. Pulido, A. Vialette-Guiraud, K. Nikovics, H. Morin, A. Hay, I.E. Johansen, M. Tsiantis, and P. Laufs. A conserved molecular framework for compound leaf development. *Science*, 322(5909):1835–1839, Dec 2008.

[5] K.J. Schmid, T.R. Sorensen, R. Stracke, O. Torjek, T. Altmann, T. Mitchell-Olds, and B. Weisshaar. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res*, 13(6A):1250–1257, Jun 2003.

[6] M. Kass, A. Witkin, and D. Terzopoulos. Snakes - Active Contour models. *International Journal of Computer Vision*, 1(4):321–331, 1987.

[7] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Symp. on Math. Statist. and Prob.*, 1967.

[8] T. M. Martinetz and K. J. Schulten. A neural-gas network learns topologies. In T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, editors, *Artificial Neural Networks*, pages 397–402. North-Holland, Amsterdam, 1991.

[9] T. Kohonen. *Self-Organizing Maps*. New York : Springer-Verlag, 1997.

[10] T. M. Martinetz, S. G. Berkovich, and K. J. Schulten. 'Neural-Gas' network for vector quantization and its application to time-series prediction. *IEEE Transactions on Neural Networks*, 4(4):558–569, July 1993.

[11] G. A. Carpenter, S. Grossberg, and D. Rosen. ART 2-A: an adaptive resonance algorithm for rapid category learning and recognition. In *Proc. IJCNN-91-Seattle International Joint Conference on Neural Networks*, pages 151–156, 1991.

[12] Juha Vesanto, Johan Himberg, Esa Alhoniemi, and Juha Parhankangas. Self-organizing map in Matlab: the SOM toolbox. In *Proceedings of the Matlab DSP Conference*, 1999.

[13] J.C. Dunn. Well separated clusters and optimal fuzzy partitions. *J.Cybern.*, 4:95–104, 1974.

[14] D.L. Davies and D.W. Bouldin. A cluster separation measure. *Trans. Pattern Anal. Machine Intell.*, 1(4):224–227, 1979.

[15] L. Hubert and J. Schultz. Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical PsychologySchn*, 29:190–241, 1976.

[16] P.J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.