# Estimating the Number of Clusters in DNA Microarray Data

**Nadia Bolshakova[1], Francisco Azuaje[2]**
[1]*Department of Computer Science, Trinity College Dublin, Ireland*
[2]*School of Computing and Mathematics, University of Ulster, Jordanstown, Northern Ireland, U.K*

Recent progress in DNA microarray technology allows measuring the simultaneous gene expression of thousands of genes under multiple experimental conditions [1]. This technology is having a significant impact on genomic and post-genomic studies. Disease diagnosis, drug discovery and toxicological research benefit from the of microarray technology.

A principal step in the analysis of gene expression data is the detection of samples or gene groups with similar expression patterns. The accurate classification of tumours is essential for a successful diagnosis and treatment of cancer. One of the problems associated with cancer tumour classification is the identification of unknown classes using gene expression profiles. Clustering is a fundamental approach to gene expression knowledge discovery. Some solutions for the systematic evaluation of the quality of the clusters have been recently proposed [2]. Moreover, the prediction of the correct number of clusters is a critical problem in unsupervised classification problems [3]. Three clustering and two validations algorithms were applied to two cancer tumour datasets. Recent studies confirm that there is no universal pattern recognition and clustering model to predict molecular profiles across different datasets. Thus, it is useful not to rely on one single clustering or validation method, but to apply a variety of approaches. Therefore, combination of these methods may be successfully used for the estimation of the number of clusters. It has been shown that these methods may support the prediction of the optimal partition and computational diagnosis. A weighed voting technique was used to improve the prediction of the number of clusters based on different data mining techniques.

The methods implemented in this research may contribute to the validation of clustering results and the estimation of the number of clusters. The results show that this estimation approach may represent an effective tool to support biomedical knowledge discovery and healthcare applications.

**References**

[1] Eisen M.B., Spellman P.T., Brown P.O., Botstein D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA (1998) 14863-14868

[2] Yeung K.Y., Haynor D.R., Ruzzo W.L.: Validating clustering for gene expression data. Bioinformatics (2001) 309-318

[3] Bolshakova N., Azuaje F.: Cluster validation techniques for genome expression data. Signal Processing (2003) 825-833