

A robust classifier for molecular tumor diagnostics

M. Laabs, F. Kleinjung, A. Malik, J. Schuchhardt
 MicroDiscovery GmbH, Marienburger Str. 1, 10405 Berlin

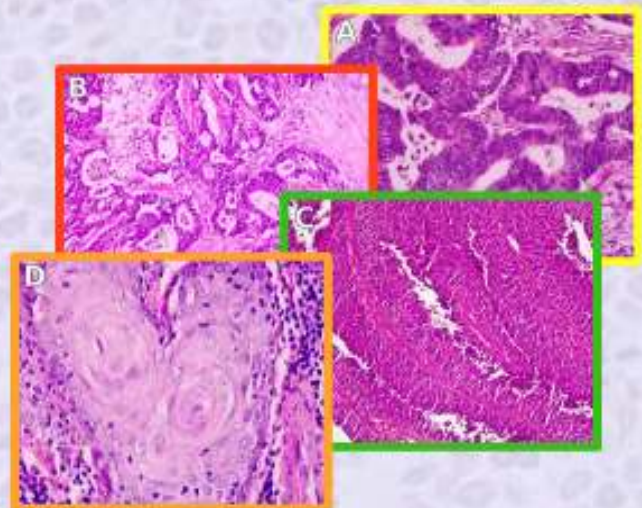


A Microarray-based Multiclass Classification System with respect to unknown classes

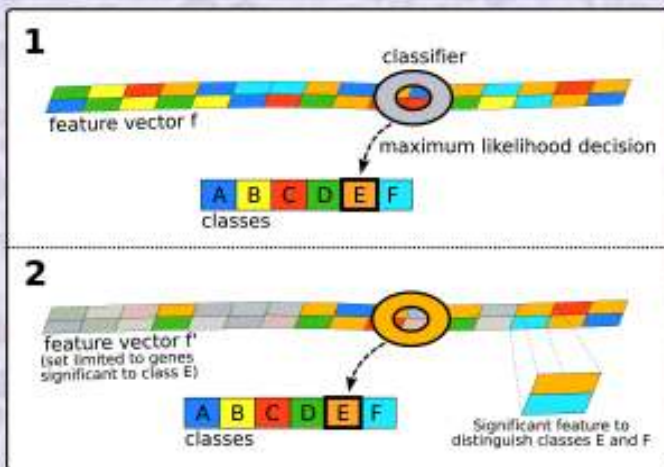
Classification of complex multi-class patterns is an important task of modern diagnostics. Precise information on the patient's disease is crucial for optimal treatment. Due to today's quality and availability of Microarray Technology, experts can rely on support by Microarray-based systems.

Often, only a subset of all existing disease subtypes is known. In this case classification systems are confronted with samples of classes alien to them. This problem receives only little attention by most current classification systems and is usually solved by the usage of thresholds, a strategy which proved suboptimal during our study.

With this in mind we present a method that aims to reduce the rate of false positives and false negatives. This is performed by a specificity test after the classification step. The specificity test assesses whether the classification is possibly caused by contributions of features that are not class specific.

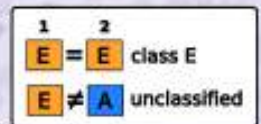


A Adenocarcinoma recti
 B Adenocarcinoma coli
 C Urothelial papillary carcinoma
 D Carcinoma planoepitheliale keratodes laryngis



MicroDiscovery Classifier

In step one, a classifier determines the class to which a sample, represented by a feature vector f , belongs, based on a maximum likelihood decision. In the second step, false positives are identified by performing a new classification with a smaller feature vector f' that only contains features specific to the class determined in step one. If both steps yield the same result the sample is assigned to the according class. Otherwise it is left unclassified.

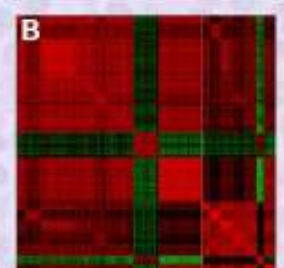
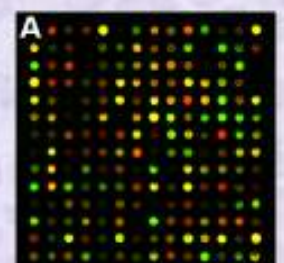
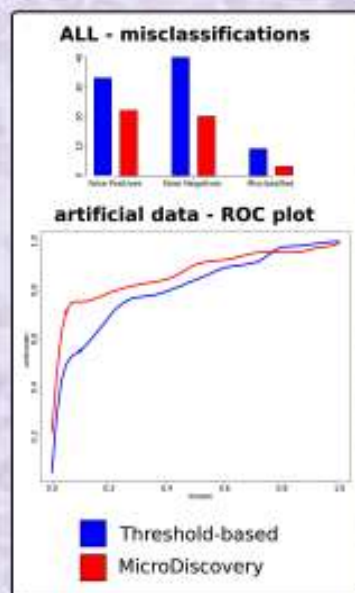


The proposed method is independent of the underlying classifier system: We successfully tested it with a Bayes Classifier as well as a Support Vector Machine.

Results

We applied our method to two datasets: An Acute Lymphoblastic Leukemia (ALL) subtype classification study (St. Jude Children's Research Hospital), and a bacteria classification pilot study on Salmonella Enterica Serovar Typhimurium. In both studies, the MicroDiscovery method resulted in fewer false positives and false negatives than the threshold-based method.

Additionally, we tested our method using an artificial dataset containing 10 different classes, 6 of which were known (trained) to the system and 4 of which were unknown. Here we were able to reduce the amount of false positives and false negatives by up to 50%.



A Microarray scan
 B Correlation plot