World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:11, No:02, 2017

Assessing Significance of Correlation with Binomial Distribution

Authors : Vijay Kumar Singh, Pooja Kushwaha, Prabhat Ranjan, Krishna Kumar Ojha, Jitendra Kumar

Abstract : Present day high-throughput genomic technologies, NGS/microarrays, are producing large volume of data that require improved analysis methods to make sense of the data. The correlation between genes and samples has been regularly used to gain insight into many biological phenomena including, but not limited to, co-expression/co-regulation, gene regulatory networks, clustering and pattern identification. However, presence of outliers and violation of assumptions underlying Pearson correlation is frequent and may distort the actual correlation between the genes and lead to spurious conclusions. Here, we report a method to measure the strength of association between genes. The method assumes that the expression values of a gene are Bernoulli random variables whose outcome depends on the sample being probed. The method considers the two genes as uncorrelated if the number of sample with same outcome for both the genes (Ns) is equal to certainly expected number (Es). The extent of correlation depends on how far Ns can deviate from the Es. The method does not assume normality for the parent population, fairly unaffected by the presence of outliers, can be applied to qualitative data and it uses the binomial distribution to assess the significance of association. At this stage, we would not claim about the superiority of the method over other existing correlation methods, but our method could be another way of calculating correlation in addition to existing methods. The method uses binomial distribution, which has not been used until yet, to assess the significance of association between two variables. We are evaluating the performance of our method on NGS/microarray data, which is noisy and pierce by the outliers, to see if our method can differentiate between spurious and actual correlation. While working with the method, it has not escaped our notice that the method could also be generalized to measure the association of more than two variables which has been proven difficult with the existing methods.

Keywords : binomial distribution, correlation, microarray, outliers, transcriptome
Conference Title : ICCSDA 2017 : International Conference on Computational Statistics and Data Analysis
Conference Location : Mumbai, India
Conference Dates : February 07-08, 2017