

Eine neue Resampling-Methode zur Untersuchung der Stabilität von Clustern bei der Genexpressionsanalyse

Gana Dresen I¹, Boes T¹, Hüsing J², Neuhäuser M¹, Jöckel KH¹

¹Institut für Medizinische Informatik, Biometrie und Epidemiologie, Universitätsklinikum Essen, Deutschland

²Koordinierungszentrum für Klinische Studien, Universitätsklinikum Heidelberg, Deutschland

irina.gana-dresen@medizin.uni-essen.de

Einleitung und Fragestellung Ein Problem bei der Interpretation der Ergebnisse von Clusteranalysen ist die Ermittlung der Verlässlichkeit der gefundenen Cluster. Statistische Methoden sind nötig, um reale Cluster von zufälligen Clustern zu unterscheiden. Neben anderen Verfahren kommen häufig Resampling-Verfahren wie der Bootstrap zur Anwendung. Das Bootstrap-Verfahren entspricht einer Ziehung mit Zurücklegen, so dass in der erzeugten Stichprobe nicht alle der ursprünglichen Beobachtungen (Gene) vorkommen. Ein Teil bleibt daher unberücksichtigt.

Das hier vorgestellte neue Verfahren beruht auf der Anwendung einer Gewichtungsmatrix, die im Gegensatz zum Bootstrap-Verfahren auf Nullelemente verzichtet und dafür nichtganzzahlige Diagonalelemente zulässt. Dadurch wird erreicht, dass die volle Dimensionalität des Raumes gewahrt bleibt, d.h. jede Beobachtung aus dem Original-Datensatz ist auch in der Resampling-Stichprobe vertreten.

Material und Methoden Zum Vergleich der neuen Methode mit dem Bootstrap-Verfahren wurde der Datensatz von Tschentscher et al. [1] verwendet, der aufgrund der Chromosomenzugehörigkeit der Gene in 24 kleinere Datensätze unterteilt wurde.

Als Grundlage zur Ziehung der Gewichtungsmatrizen diente die Lognormalverteilung. Aus den erhaltenen Resampling-Stichproben wurden mit Hilfe der hierarchischen Clusteranalyse Dendrogramme erzeugt. Die Zusammenfassung der einzelnen Dendrogramme zu einem „majority rule“ Konsensus-Baum erfolgte unter Berücksichtigung der in der Literatur angegebenen Methoden. Der verallgemeinerte Rand-Index bestimmte das Maß der Übereinstimmung zwischen dem ursprünglichen Dendrogramm und dem „majority rule“ Konsensus-Baum.

Zum Vergleich mit den Gewichtungsmatrizen wurde das Bootstrap Verfahren auf dieselben Datensätze angewendet und die erhaltenen Dendrogramme ebenfalls zu Konsensus-Bäumen zusammengefasst. Es folgte ein Vergleich dieser Konsensus-Bäume mit denen, die durch Anwendung der Gewichtungsmatrizen erzeugt wurden, wobei wieder der verallgemeinerte Rand-Index Verwendung fand.

Ergebnisse und Diskussion Im Vergleich zum Bootstrap-Verfahren sind die Konsensus-Bäume, die durch Anwendung der Gewichtungsmatrizen erhalten werden meistens differenzierter. Bei Untersuchung der Gene auf einigen Chromosomen werden mit dem Bootstrap-Verfahren und bei Verwendung der Gewichtungsmatrizen die gleichen Konsensus-Bäume erhalten. Lediglich die Häufigkeit, mit der die Cluster in den einzelnen Stichproben vorkommen ist bei Verwendung der Gewichtungsmatrizen teilweise höher.

Bei anderen Chromosomen ist mit Hilfe der Gewichtungsmatrizen eine genauere Aussage über die Clusterzugehörigkeit einiger Beobachtungen zu treffen.

Der Vorteil der Gewichtungsmatrizen gegenüber dem Bootstrap-Verfahren kommt besonders bei sehr kleinen Datensätzen zum Tragen, da es gerade hier wichtig ist, alle Beobachtungen zu berücksichtigen. Aber auch bei großen Datensätzen kann das neue Verfahren eingesetzt werden, da es aufgrund unserer bisherigen Ergebnisse auf keinen Fall schlechter als das herkömmliche Bootstrap-Verfahren ist.

Literatur

- [1] Tschentscher F, Hüsing J, Hölter T, Kruse E, Gana Dresen I, Jöckel KH et al. Tumor Classification Based on Gene Expression Profiling Shows That Uveal Melanomas with and without Monosomy 3 Represent Two Distinct Entities. *Cancer Res* 2003, 63, 2578-84.