Simple and Effective Visual Models for Gene Expression Cancer Diagnostics

Gregor Leban, Minca Mramor, Ivan Bratko und Blaz Zupa

Zusammenfassung

Seminar Data Mining in Datenbanken

Martin-Luther-Universität Halle-Wittenberg
Institut für Informatik

Joachim Bargsten

Sommersemester 2006

1 Einleitung

Der Artikel „Simple and Effective Visual Models for Gene Expression Cancer Diagnostics“ von Gregor Leban, Minca Mramor, Ivan Bratko und Blaz Zupan, erschienen im Jahr 2005, greift die Auftrennung von Krebs-Genexpressionsdaten mit einfachen 2D-Darstellungen auf. Der vorgestellte Algorithmus „VizRank“ sucht unter der extrem großen Menge möglicher Projektionen die besten raus und ermöglicht Fachleuten die Daten zu interpretieren.

Die Expressionsdaten, die dem Artikel zugrunde liegen, wurden mit Microarrays bestimmt. Microarrays sind in der Lage die Expressionswerte verschiedenster Zellen festzuhalten. Dadurch bietet sich die Möglichkeit, die Ursachen von unkontrolliertem Zellwachstum und damit die Entstehung von Krebs zu erforschen. Aber nicht nur die Ursachen, sondern auch die unterschiedlichen Krebsarten lassen sich deutlich leichter und sicherer bestimmen. Medizinisch gesehen könnte es für frühzeitige Diagnosen genutzt werden, die dann eine gezielte Therapie möglich machen.

Trotz dieses Potentials ist die Analyse nicht ohne Probleme. Die große Anzahl an Genen, bzw. Expressionsdaten bei gleichzeitig wenigen Patienten stellt hohe Ansprüche an mögliche Lösungsansätze. Viele Ansätze, z.B. Support-Vector-Machines, finden zwar zuverlässige Gen-Abhängigkeiten, doch erstrecken sich diese über hunderte von Genen und kein Experte ist in der Lage das zu interpretieren. Ziel ist es also einen einfachen, aussagekräftigen Klassifikator und eindeutige Darstellungen zu finden, die mit wenigen Genen auskommen. Als Lösung wird der VizRank-Algorithmus vorgestellt, der direkt nach guten Projektionen, bzw. Datendarstellungen, sucht. Er benutzt eine heuristische Suche um viel versprechende Projektionen zu finden und daraus die besten zu filtern. Das Ergebnis ist eine Liste mit den aussagekräftigsten Projektionen und entsprechender Wertung. Zusätzlich sind die Gene, die zu den aussagekräftigsten Projektionen gehören, meistens biologisch relevant. Sie eigenen sich besonders gut zur Krebsvorhersage.

2 Darstellung der Daten

Zur Darstellung werden die Scatterplot- und die Radviz-Projektion verwendet. Der Scatterplot ist ein einfaches xy-Koordinatensystem in dem die Expressionswerte von zwei Genen gegeneinander aufgetragen werden. Im Gegensatz dazu kann der Radviz-Graph mehr als zwei Gene gegeneinander auftragen. Die Gene sind als Fixpunkte gleichmäßig auf einem Einheitskreis verteilt. Jedes Gen „zieht“ den Datenpunkt ähnlich wie eine Feder in seine Richtung. Die Berechnung des Datenpunktes greift auf das hookesche Gesetz zurück: Die „Zugkraft“ eines Gens errechnet sich aus der Strecke zwischen Gen-Fixpunkt und dem Datenpunkt, und dem Expressionswert als Federkonstante. Der Datenpunkt liegt dort, wo die Summe aller Kräfte Null ist (vgl. Abb. 1). Damit sich jeder Punkt innerhalb des Kreises befindet, werden die Expressionswerte normalisiert. Ein wichtiger Faktor bei dieser Projektion ist die Anordnung der Gene auf dem Kreis. Es gibt Fälle, an denen zwei sehr ähnliche Gene, die die Daten gut trennen, gegenüber liegen. Die Wirkung der beiden hebt sich auf und die Projektion wird nutzlos. Platziert man die

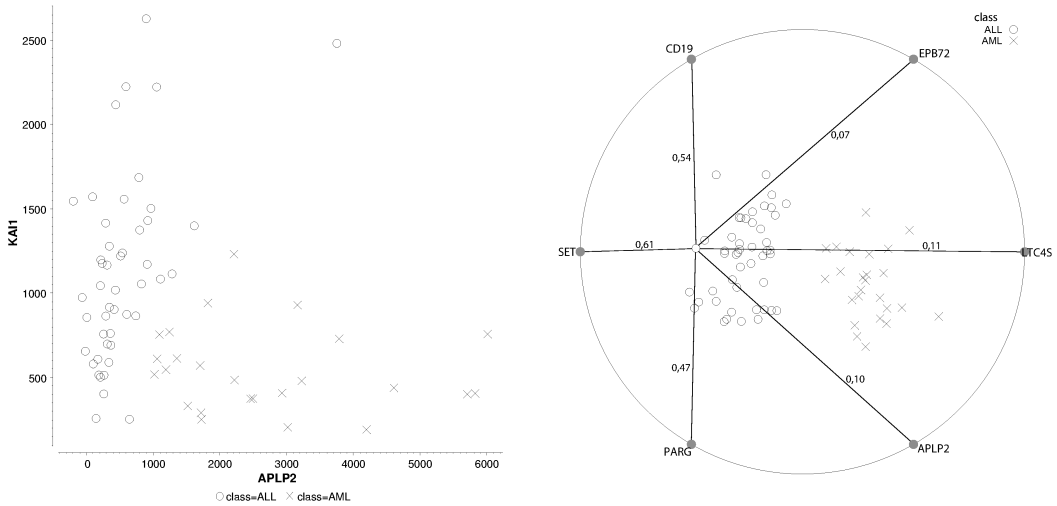

Abbildung 1: Beispiel einer Scatterplot- (links) und einer Radviz-Projektion (rechts). Die Verbindungen zu den Genen und die normalisierten Genexpressionswerte sind für einen Punkt in der Radviz-Projektion eingezeichnet.

Gene nebeneinander, ist das nicht der Fall.

3 Der VizRank Algorithmus

Die Projektionen in Abb. 1 sind sehr aussagekräftig, da die klassifizierten Daten deutlich aufgetrennt sind. Das finden solcher Projektionen übernimmt der VizRank-Algorithmus. Er berechnet eine Score für jede viel versprechende Projektion. Sind die Daten deutlich getrennt und lassen sich Regeln daraus herleiten, ist die Score hoch. Dahinter steckt eine Methode aus dem überwachten maschinellen Lernen. Aus den angelernten Daten wird die erwartete Vorhersagegenauigkeit, die Score, des Klassifikators berechnet. Eingabe sind die Punkte der grafisch repräsentierten Daten und ihre Klassenzugehörigkeit. Der Klassifikator ist ein k-nearest-neighbour-Algorithmus (k-NN) mit euklidischer Distanz. Die euklidische Distanz wurde gewählt, weil sie am ehesten dem Eindruck eines menschlichen Betrachters entspricht. Der k-NN sagt die Klassenzugehörigkeit voraus, indem die Verteilung k nächsten Nachbarn eines Punktes bestimmt wird. Ist die vorhersage der Klasse richtig, befinden sich, grafisch gesehen, nur Punkte gleicher Klasse in der Umgebung. Die Anzahl der nächsten Nachbarn ist auf $k = \sqrt{N}$ festgelegt, wobei N die Anzahl der Instanzen (Testpersonen) ist. Damit die Klassifizierung nicht so stark von N abhängt, wird zusätzlich weighted-voting benutzt. Je größer die Distanz ist, desto weniger fällt ein Punkt ins Gewicht. Das vergrößert den Einfluss nah gelegener Punkte.

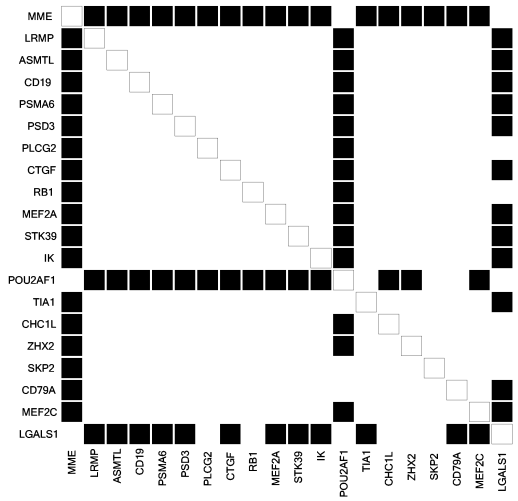

Abbildung 2: Die ersten 20 Gene der besten Scatterplot-Darstellungen eines Leukemie-Datensatzes auf der x- und y-Achse. Eine schwarze Box markiert zwei Gene, die bei den Top 500 besten Scatterplots die entsprechenden Achsen besetzen. Es wird deutlich, dass bestimmte Gene in fast allen Darstellungen vertreten sind.

4 Bewertung von Darstellungen

Ausgewertet wird das Ganze mit der leave-one-out Kreuz-Validierung. Die Score ist das Mittel der Wahrscheinlichkeit, dass der Klassifikator die richtige Klasse zuordnet. Der Anteil der falsch klassifizierten Instanzen kann hier nicht als Score genutzt werden, da z.B. bei einer Klassifikation von 100% keine Aussage über die Lage bzw. Aufteilung der Datenpunkte gemacht wird. Die Berechnung der Score ist sehr Zeitaufwendig. Wegen der exorbitanten Menge an möglichen Projektionen wird ein Verfahren gebraucht, das die Daten vorfiltert. Der VizRank Algorithmus setzt dazu eine heuristische Suche ein. Sie findet viel versprechende Projektionen deutlich schneller. Wie viel versprechend eine Projektion ist, gibt die Summe der „RELIEF-Werte“ beteiligter Gene an. Das verkleinert die Menge der zu bewertenden Projektionen drastisch und verkürzt den Suchvorgang. Eine zufällige Auswahlmethode kommt nicht in Frage, da sie deutlich schlechtere Ergebnisse liefert.

5 Experimente und Ergebnisse

Alle Ergebnisse, die der Algorithmus berechnet, haben real gemessene Werte aus klinischen Studien als Basis. Insgesamt werden acht Krebs-Genexpressionsdatensets verschiedenster Krebsarten benutzt. Bei der Auswertung fällt auf, dass in der Menge möglicher Projektionen nur ein kleiner Teil wirklich aussagekräftig ist. Ein Testlauf dauert auf einem Pentium 4 PC mit 2.4 GHz ca. zwei Stunden. Dabei wurde der VizRank Algorithmus auf 200 000 Projektionskandidaten beschränkt und die Anzahl der Gene in der

Radviz-Projektion befindet sich zwischen drei und sieben. Die Radviz-Projektion kann zwar eine willkürliche Anzahl an Genen darstellen, aber je höher die Anzahl der Gene, desto schwerer ist die Darstellung interpretierbar. Die Auswertungsdauer steigt ebenfalls signifikant an. In allen Datensets findet der VizRank-Algorithmus ausgezeichnete Darstellungen. Es könnte daher behauptet werden, dass der Algorithmus immer, auch bei zufälligen Datensets, eine ausgezeichnete Darstellung findet. Mit anderen Worten wäre der Algorithmus ein über-angepasstes (overfitted) Modell. Theoretisch gesehen ist die Wahrscheinlichkeit, dass eine zufällige Darstellung eine deutliche Klassentrennung hat, ungefähr $2.57 * 10^{-49}$ (der hier verwendete SRBCT-Datensatz hat 4 Klassen und 83 Instanzen). Praktisch gesehen hat ein zufällig permutierter Datensatz (ebenfalls SRBCT als Basis) keine einzige verwertbare Darstellung. Diese Tatsachen lassen den Schluss zu, dass in den Krebsdaten eine Regelmäßigkeit vorliegt. Die Analyse der besten Darstellung bestätigt dies. Viele Gene, die dort vertreten sind, sind auch in dem „atlas of genetic and cytogenetics in oncology and haematology“ als krebserregend oder krebsbezogen gekennzeichnet. Bezieht man auch die Darstellungen ab Platz zwei mit ein, sind manche Gene besonders häufig vertreten (vgl. Abb. 2). Diese wurden bereits von anderen Untersuchungen als spezifisch für die Krebsart deklariert. Leider sind Irrtümer nicht ausgeschlossen. Trotzdem bietet die vorgestellte Analyse einfache und effiziente Methoden Krebsarten voneinander zu unterscheiden, wobei die Darstellungen sogar von menschlicher Seite interpretierbar sind.