Ein schneller Klassifikations-Ansatz für das Screening von Zervix-Proben basierend auf einer linearen Approximation des Sammon-Mappings

Heiko Volk, Christian Münzenmayer, Matthias Grobe und Thomas Wittenberg

Fraunhofer-Institut für Integrierte Schaltungen, 91058 Erlangen
Email: vlk@iis.fraunhofer.de

Zusammenfassung. Kommt es bei einer Klassifikation auf die Verarbeitungsgeschwindigkeit an, so wird in der Regel der Polynom- gegenüber dem kNN-Klassifikator bevorzugt. Die Eigenschaft, den für die Klassifikation verantwortlichen Datensatz zu identifizieren, geht bei dem Polynomklassifikator und anderen Verfahren, wie etwa neuronalen Netzen, verloren. Die Nachvollziehbarkeit spielt aber gerade in medizinischen Anwendungen eine wichtige Rolle. Dieser Beitrag stellt einen neuen Ansatz vor, mit dem die Eigenschaften beider Klassifikatoren, die Nachvollziehbarkeit und die hohe Verarbeitungsgeschwindigkeit, kombiniert werden können. Das Verfahren beruht auf einer linearen Approximation des Sammon-Mappings. Der praktische Einsatz anhand des automatischen Zervix-Screenings zeigt die Nutzbarkeit des vorgestellten Verfahrens.

1 Problemstellung

Krebs ist eine der häufigsten Todesursachen in den Industrie-Ländern. Nach Brust- und Darmkrebs steht das Zervix-Karzinom (Gebärmutterhalskrebs) an dritter Stelle der bei Frauen auftretenden Krebsarten. Durch die ab dem 20. Lebensjahr gesetzlich vorgesehene mögliche Krebsvorsorgeuntersuchung ist die Zahl der später an Zervix-Karzinom erkrankten Fälle in den letzten Jahren stark gesunken. Die Anzahl der dabei zu untersuchenden Proben pro Jahr ist enorm. Im Rahmen eines Forschungsprojektes zur Automatisierung des Zervix-Screenings werden Bildverarbeitungs-Ansätze zur automatischen Auswertung von Zervix-Proben untersucht. Das Ergebnis einer solchen Auswertung soll als Diagnosevorschlag dem Arzt präsentiert werden.

2 Stand der Forschung

Für eine automatische Auswertung wird eine Zervixprobe unter dem Mikroskop aufgenommen und digitalisiert. Auf den Bildrohdaten werden die einzelnen Zellen segmentiert [1] und jede Zelle anschließend klassifiziert. Dies geschieht anhand einer Datenbank von Referenz-Zellen, von denen die jeweilige Klassenzugehörigkeit bekannt ist (Gold-Standard). Zur Klassifikation wird bislang der

k-Nächste-Nachbar (kNN) Klassifikator verwendet, da er die Möglichkeit bietet, die Merkmalsvektoren zu identifizieren, nach denen eine Zelle klassifiziert wurde. Dies erlaubt eine iterative Verbesserung der Trainingsstichprobe, da sich Inkonsistenzen in den Trainingsdaten somit zurückverfolgen lassen. Ebenso können dem Arzt damit die dazugehörigen Zellen der Trainings-Stichprobe graphisch präsentiert werden. Eine typische Zervix-Probe besteht aus ca. 600 digitalisierten Aufnahmen (Bildgröße 1300x1024), die in ca. 10 - 15 Minuten analysiert werden müssen. Diesen Zeitanforderungen wird der kNN, selbst in ausgedünnter und optimierter Form, bei großen Trainingsstichproben nicht gerecht. Ein Polynom-Klassifikator oder neuronale Netze, die eine schnellere Klassifikation ermöglichen, erlauben hingegen keine Kontrolle der Klassifikationsentscheidung, sind also nicht nachvollziehbar.

3 Sammon-Mapping

Das Sammon-Mapping [2] ist ein nichtlineares Mapping-Verfahren, welches zur Analyse von hochdimensionalen Daten benutzt wird. Das Hauptproblem bei der Analyse hochdimensionaler Daten ist die Erkennung von Struktur, wobei sich Struktur hier auf die geometrische Beziehung der Datenpunkte untereinander bezieht. Für gewöhnlich projiziert der Algorithmus die Daten in einen zweidimensionalen Raum. Dabei ist zu beachten, dass der Abstand zwischen zwei Punkten im 2D-Raum den korrespondierenden Abstand im hoch-dimensionalen Raum approximiert. Durch diese Randbedingung bleibt die inhärente Struktur der Datenpunkte erhalten. Mathematisch ist die Randbedingung als Kostenfunktion s_{stress} definiert:

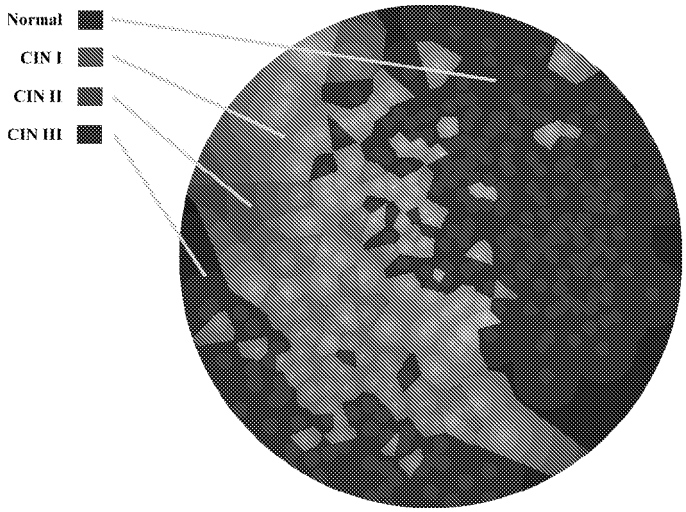
$$s_{\text{stress}} = \frac{1}{\sum\limits_{i<j} d^*_{ij}} \sum_{i<j}^{N} \frac{\left(d^*_{ij} - d_{ij}\right)^2}{d^*_{ij}} \tag{1}$$

wobei d^*_{ij} für den Abstand zweier Datenpunkte im Hochdimensionalen, d_{ij} für den Abstand im Zweidimensionalen und N für die Anzahl der Datenpunkte steht. Die Projektion stellt ein Optimierungsproblem dar, welches hier mit einem Gradientenabstiegsverfahren gelöst wird.

4 Lineare Approximation

Das nachfolgend erläuterte Verfahren zur linearen Approximation des Sammon Mappings (nachfolgend LASM genannt) kann als Hybrid-Ansatz zwischem dem kNN- und Polynom-Klassifikator betrachtet werden. In der vorliegenden Anwendung des automatischen Zervix-Screenings, beläuft sich die Dimension des Merkmalsraumes auf anfänglich 14 Merkmale. Diese werden durch das Sammon-Mapping auf einen zwei-dimensionalen Merkmalsraum projiziert. Aus den Originaldaten A und projizierten Merkmalsdaten B lässt sich ein lineares Gleichungssystem erstellen und der Vorgang des Sammon-Mappings $S(A)$ mit Hilfe einer Projektionsmatrix P schätzen.

Abb. 1. Lineare Approximation des Sammon Mappings des Zervix-Datensatzes einer Patientin. Die Darstellung der Klassenzugehörigkeit (Grün = Normale Zellen, Gelb = CIN-I Zellen, Orange = CIN-II Zellen, Rot = CIN-III Zellen, CIN = Cervical Intraepithelial Neoplasia) und zugehörigen Trainingsvektoren ist hier kombiniert dargestellt. Die Grenzen der Trainingsvektoren ergeben ein Voronoi-Diagramm, welches der Zuordnung eines unbekannten Datenpunktes dient.



$$B_{\mathrm{SM}} = S(A_{\mathrm{Original}}) \tag{2}$$

Mit Hilfe der Projektionsmatrix P lassen sich die Originaldaten in den zweidimensionalen LASM-Raum projizieren,

$$B_{\mathrm{LASM}} = P \cdot A_{\mathrm{Original}} \tag{3}$$

wobei B_{LASM} die lineare Approximation des Sammon-Mappings darstellt und allgemein $B_{\mathrm{LASM}} \neq B_{\mathrm{SM}}$ gilt.

Die projizierten Trainingsdatenpunkte B_{LASM} lassen sich als Abbildungskarte darstellen (Abbildung 1). Innerhalb dieser Karte werden die Koordinaten der zugehörigen Datenpunkte mit der jeweiligen Klasse des Datenpunktes markiert. Mit Hilfe des kNN-Klassifikators lassen sich allen übrigen Punkten der Karte der jeweilig naheliegendsten Klasse zuordnen. Während des Klassifikationsvorgangs werden unbekannte Merkmalsvektoren mittels der Abbildung P in die Karte projiziert, aus der die entsprechende Klassenzugehörigkeit ausgelesen wird. Durch Erstellen einer zweiten Karte, der sogenannten Vektorkarte, kann in gleicher Weise ein Verweis auf den jeweiligen Merkmalsvektor der Lernstichprobe gespeichert werden. Damit ist es möglich, mit einer einfachen und schnellen

Tabelle 1. Veränderung der Klassifikationsrate durch das *LASM-Verfahren im Bezug auf den kNN-Klassifikator*

Stichprobe	*Nicht optimiert*	*Optimiert*
Zervix1	[-1,6%;2,6%]	[-1,1%;0,1%]
Zervix12	[-5,4%;-0,9%]	[-4,3%;-0,2%]
Ösophagus	-30,8%	-24,4%
Vistex	-31,2%	-25,3%

Matrixmultiplikation und zwei anschließenden Speicherzugriffen die Klasse sowie den dazugehörigen Merkmalsvektor zu ermitteln.

5 Ergebnisse

Zur Evaluierung des LASM-Verfahrens wurden drei verschiedene Stichproben herangezogen, zwei medizinische - die bereits erwähnte Zervix-Zellen-Stichprobe, sowie eine Datenbank zur Untersuchung endoskopischer Aufnahmen von Barrett-Ösopahgus-Patienten [3] - sowie eine akademische, die VisTex-Stichprobe [4]. Für jede Stichprobe wurde die Gesamtklassifikationsrate für Standard- (1-NN, L2-Norm) und optimierte Einstellungen (bestes k für k-NN, bestes Abstandsmaß) des k-Nächsten-Nachbar Klassifikators ermittelt. Ebenso wurden die Raten des LASM-Verfahren berechnet und mit denen des kNN verglichen. Die Zervixzellen-Stichprobe setzte sich aus einer Probe mit einer Patientin, sowie einer Probe mit 12 verschiedenen Patientinnen zusammen. Dabei wurde jeweils ein Zwei-Klassen Problem (Gesund-Krank) und ein Mehr-Klassen Problem (Gesund, mehrere Dysplasie-Grade, Tumor) betrachtet.

Das LASM-Verfahren soll keine Steigerung der Klassifikationsrate einer Aufgabenstellung bewirken, sondern die Nachteile verschiedener Klassifikations-Verfahren durch die Kombination der jeweiligen Vorteile ausgleichen. Deshalb soll in der nachfolgenden Auswertung lediglich festgestellt werden, ob die Klassifikationsrate stabil bleibt. Die Auswirkung des LASM-Verfahrens auf die einzelnen Datensätze ist in Tabelle 1 dargestellt.

Die Änderung der Klassifikationsrate der Zervix-Zellen-Stichprobe anhand des LASM-Verfahrens mit nur einer Patientin belief sich zwischen -1,6% bis +2,6% für den Standard- und -1,1% bis +0,1% für den optimierten Fall im Vergleich zum kNN-Klassifikator. Bei der 12-Patientinnen-Probe betrugen die Raten zwischen -5,4% bis -0,9% für den Standard- und -4,3% bis -0,2% für den optimierten Fall. Die Änderungen der Klassifikationsrate bei der Ösophagus-Stichprobe betrugen -30,8% für den Standard- und -24,4% für den optimierten Fall. Ähnlich verhielten sich die Werte bei der VisTex-Stichprobe. Der Standardfall berechnete sich zu -31,2% und für den optimierten Fall ergaben sich -25,3% im Vergleich zum kNN-Klassifikator.

6 Diskussion

Für die Zervix-Zellen-Stichprobe ergibt sich eine sehr geringe Änderung im Klassifikationsverhalten des LASM-Verfahrens gegenüber dem kNN-Klassifikator, in manchen Fällen sogar eine leichte Verbesserung, was sich durch Approximationsfehler in der Projektion zugunsten des LASM-Verfahrens erklären lässt. Das etwas schlechtere Ergebnis der 12-Patientinnen-Probe zur Ein-Patientin-Probe kann aus den Unterschieden der Patientinnen untereinander erklärt werden.
Auffällig ist der starke Abfall der Klassifikationsrate bei den beiden anderen Stichproben. Es konnte gezeigt werden, dass bereits die Verringerung des Merkmalsraumes um nur eine Dimension, die Güte der Klassifikation stark beeinträchtigt. Des weiteren lässt sich durch einen Vergleich der Klassifikationsrate in Abhängigkeit der Anzahl der Dimensionen des projizierten Raumes die Dimensionalität eines Klassifikationsproblems abschätzen. So erweist sich das Zervixzellen-Problem als ein annähernd zwei-dimensionales Problem, wo hingegen die beiden anderen höher-dimensionale Probleme darstellen. Eine geringe Dimensionalität eines Problems ist jedoch nicht gleichbedeutend mit der Einfachheit deselben. Wie einfach oder komplex ein Klassifikationsproblem ist, wird hauptsächlich durch die Verteilung der Datenpunkte der einzelnen Klassen festgelegt. So kann zum Beispiel ein zwei-dimensionales Problem komplexer sein als ein drei-dimensionales. Weitere Untersuchungen zum Thema der Dimensionalität eines Klassifikationsproblem sind derzeit in Bearbeitung, wobei auch ein Vergleich mit der Hauptachsen-Transformation (PCA) vorgesehen ist.

Zusammenfassend kann gesagt werden, dass sich das LASM-Verfahren für annähernd zwei-dimensionale Problemstellungen eignet, wobei es die Vorteile zweier Klassifikationsansätze vereint. Voraussetzung für den Einsatz ist jedoch das vorherige Abschätzen der Dimensionalität eines Problems.

Literaturverzeichnis

1. Grobe M, Volk H, Münzenmayer C, Wittenberg T: Segmentierung von überlappenden Zellen in Fluoreszenz- und Durchlichtaufnahmen, Bildverarbeitung für die Medizin. Springer, Berlin, 2003.
2. Sammon J. W.: A Nonlinear Mapping for Data Structure Analysis, IEEE Transactions on Comp., Vol. C-18, No. 5, p.401-409, 1969.
3. Münzenmayer C, Mühldorfer S, Mayinger B, Volk H, Grobe M, Wittenberg T: Farbtexturbasierte optische Biopsie auf hochauflösenden endoskopischen Farbbildern des Ösophagus, Bildverarbeitung für die Medizin. Springer, Berlin, 2003.
4. MIT Media Laboratory Vision and Modelling group: Vistex vision texture database, 1995.