Einsatz von Klassifikatoren zum Lernen von Objektbeschreibungen aus hierarchisch partitionierten Bildern

Christian Thies, Marcel Schmidt-Borreda und Thomas Lehmann

Institut für Medizinische Informatik, RWTH Aachen, 52057 Aachen
Email: cthies@mi.rwth-aachen.de

Zusammenfassung. Die automatisierte explizite Extraktion von Objekten aus Bildserien, erfordert eine reproduzierbare Beschreibung der entsprechenden Bildregionen. Ein hierarchisches Partitionierungsverfahren zerlegt dazu ein Bild in seine visuell plausiblen Regionen, für die dann ein Vektor mit beschreibenden ordinalen Merkmalen berechnet wird. Objektextraktion entspricht damit der Klassifikation entsprechender Merkmalsvektoren, die vom Anwender durch markieren, trainiert werden. Da für die Klassifikation das "No-Free-Lunch-Theorem" gilt, müssen Klassifikator und Merkmalsauswahl für jede Domäne experimentell ermittelt werden. Beim Vergleich des Nearest-Neighbor Klassifikators mit dem Bayes Klassifikator mit Gaußschen Mischverteilungen und der Supportvektormaschine liefert letztere für die Handknochenextraktion aus Röntgenaufnahmen das beste Ergebnis.

1 Einleitung

In der medizinische Bildverarbeitung wächst die Menge der digital erfassten Bilder, sowie die Arten von Objekten, die in ihnen gesucht werden, in einem Maße, dass eine rein manuelle Auswertung mittelfristig nicht mehr möglich sein wird. Daher werden Verfahren immer wichtiger, mit denen a-priori unbekannte Objektbeschreibungen flexibel und effizient formulierbar sind. Hierzu ist in erster Linie eine anwenderorientierte Benutzerschnittstelle erforderlich, in der das Lernen anwendungsangepasst und mit kleinen Stichproben erfolgt.

Hierarchische Partitionierungen liefern eine vollständige Bildzerlegung in visuell plausible Regionen, die sich durch einen Merkmalsvektor beschreiben lassen [1]. Um eine solche Region als Objekt zu identifizieren, muss der regionenbeschreibende Merkmalsvektor analysiert werden, was einer Klassifikationsaufgabe entspricht. Mit diesem Ansatz lässt sich die komplexe Formalisierung von a-priori unbekannten Objekten als anwenderorientiertes "Point & Click"- Training realisieren. Problem dabei ist die Auswahl des geeigneten Klassifikators und dessen Integration in einen Workflow.

In der Literatur existieren zahlreiche Klassifikatoren, sowie Methoden zur Merkmalsreduktion, Sampling der Daten und Parametersuche [2]. Ferner gibt es Verfahren, die Bilder vollständig und hierarchisch in die visuell plausiblen

Regionen aller Skalen partitionieren [1]. Die Auswahl des Klassifikators hängt allerdings von der Struktur des verwendeten Merkmalsraums ab und lässt sich nach dem "No-Free-LunchTheorem" nicht generisch beantworten, sondern muss experimentell ermittelt werden [2].

Die mit dem hier vorgestellten Framework durchgeführten Experimente dienen dazu, die Abhängigkeit der Klassifikationsergebnisse vom Merkmalsraum, den verwendeten Parametern und dem gewählten Klassifikationsparadigma für hierarchisch partitionierte Bilder zu verstehen. Auf diese Weise wird untersucht, ob es möglich ist, Anwenderwissen durch simples Markieren relevanter Beispielregionen ohne formale Interaktion zu lernen, und dieses Wissen automatisiert zu reproduzieren. Dabei werden reale Daten der klinischen Routine verwendet.

2 Klassifikatoren für die Objektsuche

Als Werkzeug dient ein Framework, das den Datenfluss der Klassifikation modelliert und so effiziente Experimente ermöglicht. Zur Merkmalsreduktion stehen die Lineare Diskriminanz Analyse (LDA), die Hauptkomponenten Analyse (PCA) sowie das auf- und absteigende Greedy-Verfahren zur Verfügung. Als Klassifikatoren wurden die Supportvektormaschine (SVM), der Bayes Klassifikator mit Gaußschen Mischverteilungen (GMM) und das k-Nearest-Neighbor-Ähnlichkeitsmaß (KNN) gewählt. SVM und GMM sind modellbasiert, wobei die SVM versucht, strukturelles und empirisches Risiko zu minimieren, während GMM zwar theoretisch durch empirische Risikominimierung immer das beste Resultat liefert, dieses aber mit Overfitting einhergeht. Im Vergleich dazu wird der KNN-Klassifikator als modellfreier Ansatz verwendet. Die Tests werden mit unterschiedlich gesampelten Test- und Trainingsdaten durchgeführt. Dazu steht das Bootstrapping zur Verfügung. Beim GMM kann zusätzlich Varianz-Pooling verwendet werden. Aus Anwendersicht ist bei der Klassifikation die Zahl der korrekt identifizierten Objekte entscheidend. Dies sowohl im Verhältnis zur Gesamtzahl aller Rückgaben (Precision) als auch in Bezug zu allen im Datensatz vorhandenen Werten (Recall). Die Accuracy als Maß der Klassifikationsgüte ist für die Objektsuche von untergeordneter Bedeutung. Daher wird das F-Measure als harmonisches Mittel aus Precision und Recall als Qualitätsmaß verwendet [3].

$$\text{F-Measure} = \frac{2 \cdot \text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{1}$$

Ziel der Experimente ist also die Bestimmung des Klassifikators, der ein maximales F-Measure (Gl. 1) für die gegebene Ground-Truth liefert.

3 Experimente

Die Experimente werden auf 105 zufällig ausgewählten Handradiographien aus der klinischen Routine durchgeführt. Diese Bilder werden z.B. für die vergleichende Auswertung von Handknochen zur Maturitätsbestimmung bei Heranwachsenden eingesetzt. In dieser Anwendung würde eine automatisierte Extraktion der Knochen dem befundenden Radiologen reproduzierbare Abmessungen

der Objekte liefern. Für die Experimente wurden die Mittelhandknochen als Referenzobjekte gewählt.

Die Bilder werden auf eine maximale Kantendimension von 256 Pixeln unter Beibehaltung der Aspect-Ratio verkleinert. In dieser Größe sind die gesuchten Handknochen immer noch erkenn- und vermessbar, wobei der Raum der zu durchsuchenden Daten beschränkt bleibt. Die Partitionierung erfolgt mittels eines hierarchischen Bereichswachstumsverfahrens [1], das die Verschmelzung von Pixeln über Regionen hin zum gesamten Bild protokolliert. Dabei werden für alle Regionen 38 beschreibende Merkmale (Rundheit, Größe, Grauwertvarianz,...) berechnet. Für jedes Bild ergeben sich ca. 2.500 Regionen, was wiederum für die gesamte Bildmenge einen 43-dimensionalen Merkmalsraum mit ca. 265.000 Elementen ergibt. Als Ground-Truth wurden manuell 372 Vektoren bestimmt, die die erkennbaren Mittelhandknochen im Datenraum repräsentieren. Alle anderen Vektoren fallen in die Rückweisungsklasse.

3.1 Merkmalsauswahl

Im ersten Schritt wird die Abhängigkeit des Ergebnisses von der Merkmalsauswahl und damit ein optimaler Merkmalssatz bestimmt. Die Tests erfolgen dabei über eine zufällige Auswahl von 10 Trainings- und 53 Testbilder aus der Ground-Truth von 105 Bildern, also etwa 130.000 Merkmalsvektoren und 50 Trainigsobjekte.

Greedy-Verfahren. Für alle drei Klassifikatoren wird jeweils das auf- und absteigende Greedy-Verfahren, zur Merkmalsauswahl angewandt. Die Parametersuche für die SVM erfolgt dabei mittels Grid-Suche und Gaußschem Kernel (RBF), für den GMM-Klassifikator wird ein Cluster sowie Varianz-Pooling verwendet. Der KNN wird für den ersten Nachbarn (1-NN) mit minimaler Euklidischer Norm (L2), Mahalanobis-Distanz (MAH) sowie Cosinusnorm (COS) untersucht.

Korrelationsanalyse. Zum Vergleich mit den beiden Greedy-Verfahren werden die PCA und die LDA für eine bis 38 Zieldimensionen jeweils für alle drei Klassifikatoren getestet.

3.2 Anwendungsszenario

Im zweiten Schritt wird das Anwendungsszenario simuliert. Es entspricht dem Markieren von gesuchten Objekten in einer kleinen und zufälligen Auswahl von Bildern durch den Anwender. Mit den markierten Vektoren wird dann der Klassifikator trainiert, und auf die verbleibende Testmenge angewandt. Es werden 5, 10, 20 und 40 Mittelhandknochen gewählt, was 1,2,4, bzw. 8 aus 105 Bildern zum Training entspricht. Um die statistische Verteilung der Auswahl zu simulieren, werden jeweils 20 Samples mittels Bootstrapping gezogen.

Tabelle 1. F-Measure, Precision und Recall für das aufsteigende Greedy-Verfahren im Vergleich der untersuchten Klassifikatoren

Klassifikator	# Merkmale	Recall	Precision	F-Measure
1-NN, L2	7	0,36	0,28	0,31
1-NN, COS	12	0,52	0,28	0,36
1-NN, MAH	4	0,36	0,28	0,32
GMM	12	0,58	0,37	0,45
SVM	14	0,58	0,67	0,62

4 Ergebnisse

4.1 Merkmalsauswahl

Greedy-Verfahren aufsteigend. Bei der Berechnung der Merkmalsreduktion mit dem aufsteigenden Greedy-Verfahren war die SVM mit einem F-Measure von 0.62 und 14 Merkmalen der Klassifikator mit der höchsten erzielbaren Genauigkeit während der modellfreie 1-NN Klassifikator mit der L2-Norm das schlechteste Ergebnis liefert(Tab. 1).Von den 38 verfügbaren Merkmalen wurden maximal 14 ausgewählt. Precision und Recall variieren für die einzelnen Klassifikatoren.
Greedy-Verfahren absteigend. Mit dem absteigenden Greedy-Verfahren lieferte ebenfalls die SVM bei 32 Merkmalen das beste Ergebnis allerdings nur mit einem F-Measure von 0.46. Für den 1-NN ergaben sich mit COS bei 12 Merkmalen ein F-Measure von 0.36.
PCA. Die zum Vergleich durchgeführet PCA lieferte mit einem F-Measure von 0.21 ihr bestes Ergebnis bei einer Zieldimension von 26 für GMM. Dabei verbessert sich das Ergebnis mit wachsender Zieldimension.
LDA. Bei der LDA liefert 1-NN mit der L2 Norm das beste Ergebnis mit einem F-Measure von 0.31 bei einer Zieldimension von 23. Dabei variieren die Werte für die Zieldimensionen von 1 bis 38 nicht monoton zwischen 0.2 und 0.3.

4.2 Anwendungsszenario

Die Simulation des Anwendungsszenarios liefert für die SVM bei 20 unterschiedlichen Trainings-Samples von jeweils 40 Knochen ein bestes mittleres F-Measure von 0.42 (Tab. 2). Die angegebenen Werte sind dabei jeweils eine Mittelung aus 20 Einzelklassifikationen. Precision und Recall weisen für die SVM und den GMM Klassifikator hohe Differenzen auf, aber man erkennt einen monotonen Anstieg mit wachsender Anzahl Trainingsobjekte Ein Wilcoxon-Test auf dem 5% Niveau bestätigt außerdem das signifikant bessere Abschneiden der SVM über die 20 Samples im Vergleich zu 1-NN und GMM. Für den Grenzwert von 80 Trainingsvektoren ergeben sich keine höheren Werte.

5 Diskussion

Die SVM liefert für die gestellte Aufgabe das beste Ergebnis hierbei dient die Kombination aus hoher Generalisierungsfähigkeit bei gleichzeitiger Minimierung

Tabelle 2. Ergebnisse der Klassifikatoren bei wachsender Anzahl von Trainigsobjekten.

# Trainingobjekte	Klassifikator	Recall	Precision	F-Measure
5 Knochen	1-NN L2-Norm	0,25	0,26	0,25
	GMM (1)	0,28	0,26	0,25
	SVM	0,26	0,29	0,25
10 Knochen	1-NN L2-Norm	0,27	0,28	0,28
	GMM (1)	0,50	0,22	0,30
	SVM	0,24	0,44	0,28
20 Knochen	1-NN L2-Norm	0,29	0,29	0,29
	GMM (1)	0,30	0,30	0,30
	SVM	0,35	0,43	0,37
40 Knochen	1-NN, L2-Norm	0,32	0,33	0,33
	GMM (1)	0,60	0,28	0,38
	SVM	0,38	0,52	0,42

des empirischen Risikos dem Ausgleich der extremen Klassenschieflage zwischen Objekt- und Rückweisungsklasse. Die Greedy-Heuristik liefert die Merkmalsauswahl, die zum besten Ergebnis führt. Dies liegt an der nichtlinearen Korrelation der Merkmalsverteilungen, die von LDA und PCA nicht modelliert wird. Diese Beobachtungen lassen sich jedoch nach dem "No-Free-Lunch-Theorem" nicht ohne weiteres verallgemeinern. Für andere Aufgabenstellungen, d.h. Merkmalsräume, müssen die Experimente entsprechend wiederholt werden, wozu das vorgestellte Framework ein nachvollziehbares Werkzeug bietet.

Die Erkennungsrate für die vorgestellte Anwendung der lokalen Bildanalyse ist in jedem Fall verbesserungsfähig, sie liegt allerdings in einer Größenordnung wie sie auch aktuelle und von der Komplexität vergleichbare Ansätze zum globalen inhaltsbasierten Bilddatenbankzugriff liefern [4]. Die Experimente zeigen, das ein "Point & Click"-Training zur Beschreibung und Extraktion a-priori unbekannter Objekte mittels Klassifikation prinzipiell möglich ist. Um die Genauigkeit zu verbessern, müssen die verwendeten Merkmale und ihre Variationen jedoch weitergehend untersucht werden.

Literaturverzeichnis

1. Beier D, Thies C, Güld MO, Fischer B, Kohnen M, Lehmann TM. Ein lokaladaptives Ähnlichkeitsmaß als Kriterium der hierarchischen Regionenverschmelzung. In: Procs BVM; 2004. p. 100–4.
2. Duda RO, Hart PF, Stork DG. Pattern Classification. 2nd edition, Wiley Interscience, New York; 2001.
3. van Rijsbergen CJ. Information Retrieval. 2nd edition, Butterworths, London; 1979.
4. Clough P, Müller H, Sanderson M. The CLEF 2004 Cross-Language Image Retrieval Track. LNCS 2005;3491:597–613.