Automatische Detektion von Trockenstress bei Tabakpflanzen mittels Machine-Learning-Verfahren

Michael Siebers[1], Franz Uhrmann[2], Oliver Scholz[2], Christoph Stocker, Ute Schmid[1]

Abstract: Dieser Beitrag befasst sich mit der Klassifikation der Vitalität von Pflanzen durch Machine-Learning-Verfahren am Beispiel von Trockenstress bei Tabak (*Nicotiana tabacum*). Wir zeigen, dass Machine-Learning-Verfahren die menschliche Unterscheidung von gesunden und gestressten Pflanzen durch einen Experten nachbilden können und zudem, dass eine frühzeitige Erkennung von Pflanzenstress möglich ist, indem eine dritte Klasse für mäßig gestresste Pflanzen eingeführt wird. Zur Klassifikation werden Entscheidungsbaumverfahren, Support Vector Machine, künstliche Neuronale Netze und Lineare Regression verglichen. Im Beitrag wird schwerpunktmäßig die Auswahl der Merkmale beschrieben, die für eine zuverlässige Klassifikation notwendig sind. Da die Experteneinschätzung weniger auf Einzelkriterien als vielmehr auf dem Gesamteindruck des Pflanzenphänotyps basiert, stellt sich die Frage, welche relevanten Merkmale ein automatisches Diagnose-System berücksichtigen muss. Es hat sich herausgestellt, dass neben blattspezifischen Merkmalen auch Merkmale, die sich auf die Gesamtpflanze beziehen, für die Klassifikation relevant sind.

Keywords: Phänotypisierung, Trockenstress, Machine-Learning, Klassifikation, Merkmale

1 Einleitung

Beim „Molecular Farming“ werden Pflanzen zur Produktion von pharmazeutischen Stoffen in modernen Gewächshäusern unter kontrollierten Umgebungsparametern wie Temperatur, Wasserzufuhr oder Beleuchtung aufgezogen. Für eine vollautomatische Regelung des Gewächshaussystems fehlt jedoch bislang die Rückkopplung der Pflanzenvitalität: Beispielsweise ist die automatische Detektion von Pflanzenstress erforderlich, um die Umgebungsparameter entsprechend anzupassen oder betroffene Pflanzen aus dem System zu entfernen. Daher untersuchen wir, welches Machine-Learning-Verfahren geeignet ist, das Stresslevel von Pflanzen zu bestimmen und welche Merkmale hierfür erfasst werden müssen. Diese Arbeit basiert auf der Masterarbeit von Herrn Stocker [St13a]. Erste Ergebnisse wurden in [St13b] veröffentlicht.

Im folgenden Kapitel wird das Vorgehen zur Klassifikation der Pflanzenvitalität vorgestellt. In Kapitel 3 präsentieren wir eine Studie zur Evaluation unseres Vorgehens und Merkmale die zu einer hohen Klassifikationsgüte führen. Die Arbeit schließt mit einem kurzen Ausblick auf offene Forschungsfragen.

[1] Otto-Friedrich Universität Bamberg, Professur für Angewandte Informatik, insbes. Kognitive Systeme, An der Weberei 5, 96047 Bamberg, {vorname.nachname}@uni-bamberg.de
[2] Fraunhofer-Institut für Integrierte Schaltungen IIS, Abt. Berührungslose Mess- und Prüfsysteme (BMP), Flugplatzstr. 75, 90768 Fürth, {vorname.nachname}@iis.fraunhofer.de

2 Vorgehen

In dieser Arbeit werden verschiedene Machine-Learning-Verfahren miteinander verglichen: C4.5 zum Lernen von Entscheidungsbäumen, Backpropagation zum Lernen Neuronaler Netze, Klassifikation durch lineare Regression und Support Vector Machines (SVM). Alle vier Verfahren ordnen einer Pflanze, repräsentiert als Vektor von Merkmalen, eine Klasse, z.B. *gestresst* zu. Die konkreten Zuordnungsfunktionen werden anhand von vorklassifizierten Pflanzen gelernt. Eine detaillierte Beschreibung der Lernverfahren bieten [Mi97] und [HK05].

Evaluiert werden die Lernverfahren anhand der Accuracy, also dem Prozentsatz korrekter Klassifikationen innerhalb eines Testdatensatzes. Um eine aussagekräftige Evaluation zu bekommen wurde eine 3-fold Cross-Validation durchgeführt. Hierbei werden alle vorhandenen Daten in drei gleich große Teile aufgeteilt. Anschließend wird drei Mal aus zwei Teilen gelernt und auf einem Teil evaluiert. Dabei wird jeder Teil genau einmal zur Evaluation herangezogen. Im folgenden Abschnitt wird berichtet, welche Merkmalen erfasst wurden und wie diese gemessen wurden. Um die gelernten Zuordnungsfunktionen möglichst einfach zu halten, wurde nur ein Teil der erfassten Merkmale zur Klassifikation verwendet. Das Verfahren zur Auswahl der Merkmale wird nach dem Abschnitt zu deren Erfassung vorgestellt.

2.1 Merkmalserfassung

Die Erhebung der Merkmale ist ein dreistufiger Prozess. Zuerst wird die Pflanze vermessen. Anschließend werden die einzelnen Blätter der Pflanze als 3D-Modell rekonstruiert. Schließlich werden die Blattwerte auf Pflanzenebene aggregiert. Zur Vermessung der Pflanze wird das Lichtschnitt-Verfahren eingesetzt. In diesem Verfahren wird Laser-Licht auf die Pflanze projiziert und dessen Reflexion mit mehreren Kameras aufgenommen. Während der Messung wird die Pflanze um 360 Grad gedreht und die Reflexionspunkte verfolgt. Hierdurch erhält man eine 3D-Punktwolke, welche die Oberfläche der Pflanze darstellt.

Anschließend wird die Punktwolke mittels eines Clusteringverfahrens aufgeteilt, so dass jedes Cluster einem Blatt der Pflanze entspricht. Auf die Punktwolke jedes Blattes wird ein parametrisches Modell eines Blattes angepasst [Uh13], wobei jeder Modellparameter einer geometrischen Eigenschaft des Blattes entspricht, z.B. dessen Länge. Die Rundung und Welligkeit des Blattes werden durch die durchschnittliche Krümmung des Verlauf der Mittelrippe und des Blattrandes und deren Standardabweichungen repräsentiert. Zusätzlich zu den statischen Merkmalen jedes Blattes werden auch die Differenzen der Merkmale zur vorgehenden Messung erhoben.

Auf Pflanzenebene wird die Gesamtheit der Blätter durch die Mittelwerte der einzelnen Blattmerkmale repräsentiert. Zusätzlich wurden das Alter der Pflanze, die Gesamthöhe der Pflanze und die Gesamtfläche der Blätter in den Datensatz aufgenommen. Somit wird jede Pflanze durch insgesamt 28 Merkmale dargestellt.

2.2 Merkmalsselektion

Zur Auswahl von zur Klassifikation geeigneten Merkmalen wurde das *Forward Selection* Verfahren angewandt. Initial nimmt das Verfahren an, dass keine Merkmale zur Klassifikation benötigt werden. Anschließend werden sukzessive weitere Merkmale der Auswahl hinzufügt. Hierbei wird das nächste Attribut so gewählt, dass hierdurch die Klassifikationsgüte maximal steigt. Sollte kein weiteres Merkmal die Klassifikationsgüte erhöhen endet der Algorithmus.

3 Untersuchung

Um das vorgeschlagene Verfahren zu testen, wurden 50 Tabakpflanzen in unterschiedlichen Stressstadien gemessen. Die Pflanzen wurden hydroponisch in Steinwolleblöcken gezogen und in einem Phytotron kultiviert. Die Pflanzen wurden in 10er Gruppen im Abstand von jeweils einer Woche angesät. Zum Zeitpunkt der Messungen waren die Pflanzen zwischen drei und zehn Wochen alt.

Jede Pflanze wurde zweimal täglich mittels des Lichtschnitt-Verfahrens vermessen. Bei der initialen Messung waren alle Pflanzen ausreichend mit Wasser versorgt. Anschließend wurden zwei Pflanzen je Altersgruppe von der Wasserversorgung getrennt. Nach jeweils drei weiteren Messungen wurden weitere zwei Pflanzen je Altersgruppe von der Wasserversorgung abgeschnitten.

Die Bewertung der Pflanzenvitalität erfolgte durch einen Biologen anhand von Bildern die während der Messung der Pflanze gemacht wurden. Die Pflanzen wurden in die Kategorien *ungestresst*, *mäßig gestresst* und *gestresst* eingeteilt. Um die Objektivität der Bewertung zu gewährleisten wurden die Bilder in zufälliger Reihenfolge und ohne Information zur Bewässerung präsentiert.

Für eine erste Auswertung wurden lediglich *ungestresste* und *gestresste* Pflanzen betrachtet. Alle vier Lernverfahren erreichten Genauigkeiten von über 90% (Tabelle 1). Die lineare Regression schnitt mit 99,85% am besten ab. Diese Klassifikationsgüte konnte mit lediglich 6 Attributen erreicht werden: Alter der Pflanze, Standardabweichung der Welligkeit, Mittlerer Abstand zwischen Modell und Punktwolke, Durchschnittliche und Gesamte Blattfläche, Neigung des Blattes Richtung Boden und dem Rotationswinkel der rechten Blattseite um die Mittelachse des Blattes.

Als zweiten Auswertungsschritt sollten nun auch mäßig gestresste Pflanzen klassifiziert werden. Die erhöhte Schwierigkeit dieser Aufgabe zeigt sich auch in der Güte der Klassifikation. Die Accuracy sank auf ca. 75%, lediglich die SVM konnte eine Accuracy von gut 80% erreichen (Tabelle 1). Für diese Aufgabe genügten der SVM 5 Attribute: Standardabweichung der Welligkeit, Durchschnittlicher Höhenunterschied zwischen Blattrand und Mittelrippe, Neigung des Blattes Richtung Boden (statisch und Differenz zur vorherigen Messung) und Seitenverhältnis des Blattes.

	C4.5	Lineare Regression	Neuronales Netz	SVM
binär	92,17%	99,85%	96,98%	97,90%
tertiär	75,59%	76,61%	73,68%	80,29%

Tab. 1: Accuracy für binäre Klassifikation (*ungestresst*/*gestresst*) und tertiäre Klassifikation (*ungestresst*/*mäßig gestresst*/*gestresst*)

Die Krümmung und die Welligkeit der Blätter scheinen wichtige Indikatoren für den Trockenstress von Tabakpflanzen zu sein. Für beinahe jedes Lernverfahren wurde eines der entsprechenden Merkmale für die finale Klassifikation gewählt.

4 Diskussion und Ausblick

Wir haben ein Verfahren vorgestellt, dass es erlaubt, das Stresslevel einer Tabakpflanze zu klassifizieren. Hierbei wurde eine Genauigkeit von fast 100% erreicht, wenn lediglich *gestresste* von *ungestressten* Pflanzen unterschieden werden sollten. Bei Hinzunahme von *mäßig gestressten* Pflanzen sinkt die Genauigkeit auf ca. 80%.

Jede Pflanze wurde durch globale Pflanzenmerkmale und Merkmale ihrer Blätter repräsentiert. Hierbei wurde der Durchschnitt der Blattmerkmale erhoben. Ein weiterführender Ansatz ist, jedes Blatt einzeln zu Betrachten und die Aggregation erst nach der Klassifikation vorzunehmen. Ein weiterer Forschungsaspekt ist die Rückmeldung der Zuordnungsfunktionen an Domainexperten. Dies ermöglicht eine genauere Auseinandersetzung mit den gewählten Merkmalen und erlaubt die Zuordnungen mittels Expertenwissen anzupassen.

Literaturverzeichnis

[HK05] Han, J.; Kamber, M.: Data mining: concepts and techniques. Kaufmann, San Francisco, 2005.

[Mi97] Mitchell, T.: Machine Learning. McGraw-Hill International Editions, 1997.

[St13a] Stocker, C.: A model-based prediction of plant growth: Drought stress level classification of tobacco plants. Masterarbeit, Universität Bamberg, 2013.

[St13b] Stocker, C. et al.: A Machine Learning Approach to Drought Stress Level Classification of Tobacco Plants. In (Henrich, A.; Sperker, H.-C. Hrsg.): Lernen, Wissen & Adaptivität, Workshop Proceedings, Bamberg 2013. Media Informatics Group, University of Bamberg, S. 163–167, 2013

[Uh13] Uhrmann, F. et al.: A model-based approach to extract leaf features from 3d scans. In: Proceedings of the 7th International Conference on Functional-Structural Plant Models, Saariselkä, Finland, 2013.