Hochschule für Angewandte Wissenschaften Hamburg
Hamburg University of Applied Sciences

Bachelorarbeit

Tom Klonikowski

Ein entscheidungsunterstützendes System zur Bestimmung von Mikroorganismen unter Einsatz von Methoden der Künstlichen Intelligenz

Fakultät Technik und Informatik
Studiendepartment Informatik

Faculty of Engineering and Computer Science
Department of Computer Science

Tom Klonikowski

Ein entscheidungsunterstützendes System zur Bestimmung von Mikroorganismen unter Einsatz von Methoden der Künstlichen Intelligenz

Bachelorarbeit eingereicht im Rahmen der Bachelorprüfung

im Studiengang Bachelor of Science Angewandte Informatik
am Department Informatik
der Fakultät Technik und Informatik
der Hochschule für Angewandte Wissenschaften Hamburg

Betreuender Prüfer: Prof. Dr. Michael Neitzke
Zweitgutachter: Prof. Dr. Olaf Zukunft

Eingereicht am: 23. August 2012

Tom Klonikowski

Thema der Arbeit
Ein entscheidungsunterstützendes System zur Bestimmung von Mikroorganismen unter Einsatz von Methoden der Künstlichen Intelligenz

Stichworte
Entscheidungsunterstützung, Bestimmung, Klassifikation, Symbolische Datenanalyse, Mikroorganismen, Künstliche Intelligenz

Kurzzusammenfassung
Beschreibungen von Mikroorganismen liegen in der Regel in aggregierter Form, bezogen auf Taxa, vor und haben durch die Abbildungen der Variationen innerhalb eines Taxons ein komplexe Struktur. Dadurch ergeben sich besondere Anforderungen an die Problemlösungsmethodik für das Klassifikationsproblem der Bestimmung von Mikroorganismen. In dieser Bachelorarbeit werden klassische Problemlösungsmethoden wissensbasierter Systeme unter Verwendung symbolischer Daten untersucht, um darauf aufbauend Algorithmen für eine Softwarekomponente zu entwerfen. Die Softwarekomponente soll im Rahmen einer Software zur Gewässergüteklassifikation auf Basis des Saprobiensystems eingesetzt werden.
Tom Klonikowski

Title of the paper
A Decision Support System for taxonomic identification of microorganisms by means of Artifical Intelligence

Keywords
Decision Support, Taxonomic Identification, Classification, Symbolic Data Analysis, Microorganisms, Artificial Intelligence

Abstract
Descriptions of microorganisms are usually available in an aggregated form, related to taxa. In order to take account of the variation within a taxa, these descriptions have a complex structure, putting special demands on methodology used to solve the classification problem made up by the identification of a microorganism. This document analyzes standard problem-solving methods of knowledge based systems considering symbolic data and proposes a design for a software component.

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Listings

1. Einleitung

1.1. Motivation

Für eine Software zur Berechnung des Saprobienindex nach DIN 38410-1:2004-10[1] soll eine interaktive Komponente entwickelt werden, die die für dieses Verfahren notwendige Bestimmung von Mikroorganismen anhand mikroskopischer Bilder unterstützt.
Der Saprobienindex gibt Aufschluss über die Belastung eines Fließgewässers. Das zur Berechnung eingesetzte Verfahren ist grundsätzlich mit Hilfe eines geeigneten Mikroskops durchführbar und daher auch dezentral einsetzbar. Es erscheint daher wünschenswert, die Bestimmung der Mikroorganismen so zu vereinfachen, dass die notwendige Expertise des Benutzers minimiert wird und das Verfahren so idealerweise auch für interessierte Laien anwendbar wird.

1.2. Ziel

In dieser Arbeit wird untersucht, in welcher Form das Wissen über Mikroorganismen aufzubereiten ist und welche Lösungsmethoden für das Bestimmungsproblem bekannt sind. Es werden die Anforderungen an eine Softwarekomponente für die einfache und nutzerfreundliche Bestimmung beschrieben. Außerdem sollen effiziente Algorithmen für die Problemlösungsmethoden gefunden und implementiert werden.

1.3. Aufbau der Arbeit

Im zweiten Abschnitt wird das Problem der Bestimmung von Mikroorganismen und die gebräuchlichen Verfahren dazu beschrieben und die Grundlagen der Anwendungsdomäne umrissen.

Die informationstheoretischen Grundlagen werden im dritten Kapitel dargelegt. Dort wird zunächst der Bereich der wissensbasierten Systeme umrissen. Die Einordnung als Klassifi-

[1] DIN384101 (2004).

kationsproblem wird hergeleitet und bekannte Problemlösungsmethoden dargelegt und diskutiert. Nachfolgend wird ein kurzer Überblick über die symbolische Datenanalyse, die die theoretische Grundlage für das Datenmodell bietet, gegeben und der Structured Descriptive Data-Standard als Datenformat vorgestellt.

Die Beschreibung der Anforderungen an die Softwarekomponente erfolgt in Kapitel IV. Im fünften Abschnitt werden Algorithmen für die Testauswahl und die Bewertung von Lösungen diskutiert.

Das Saprobiensystem sowie die Methode der Berechnung des Saprobienindex werden im Anhang erläutert.

2. Problembeschreibung

2.1. Grundlagen der Anwendungsdomäne

2.1.1. Einsatzzweck

Die Komponente zur Bestimmung von Mikroorganismen mithilfe mikroskopischer Bilder soll im Rahmen einer Software für das Monitoring der Belastung von Fließgewässern auf Basis des Saprobiensystems eingesetzt werden.

Die genaue Bestimmung der in Proben enthaltenen Saprobien ist wesentliche Voraussetzung für die Berechnung des Saprobienindex nach DIN 38410-1:2004-10. Für dieses Verfahren ist in der Regel die Bestimmung bis zur Ebene der Art[1] notwendig. Ausgangsbasis für die Bestimmung sind Bilder, die nach geeigneter Präparation der Probe mit einem Lichtmikroskop erzeugt werden.

2.1.2. Bestimmung von Organismen

Die Klassifizierung eines einzelnen Individuums wird als Bestimmung bezeichnet. Dabei vergleicht der Anwender die beobachteten Merkmale eines zu untersuchenden Individuums mit einer Wissensbasis, die den Taxa bestimmte Merkmalsausprägungen zuweist. Aus Übereinstimmungen kann er so auf die Zugehörigkeit des Organismus zu einem Taxon schliessen. Diese Wissensbasis kann sich aus der Erfahrung des Anwenders ergeben, aus Fachliteratur bezogen werden oder in einer Datenbank vorliegen.

Für die Bestimmung von Organismen sind verschiedene Verfahren entwickelt worden. Dabei kann zwischen „Interaktiven Verfahren“ und „Automatischen Verfahren“ unterschieden werden (Hagedorn (2007)). Interaktive Verfahren beziehen Informationen über die Merkmalsausprägungen des zu untersuchenden Individuums vom Nutzer. Automatische Verfahren können vollständig maschinell durchgeführt werden. Sie basieren z. B. auf Bilderkennung oder auf der Erkennung von Mustern der DNA[2]. Bisher sind jedoch hauptsächlich Mischformen ge-

[1] Für die Bezeichnung der taxonomischen Ebene „Art“ wird oft auch das Synonym „Spezies“ verwendet.
[2] Miller u. a. (2009).

bräuchlich, die den Anwender bei Bestimmung unterstützen, da die automatischen Verfahren nicht immer eindeutige Ergebnisse liefern.

Bestimmungsschlüssel. Ein seit langem gebräuchliches Verfahren sind Bestimmungsschlüssel genannte Entscheidungsbäume. Die inneren Knoten dieser Bäume bilden Tests auf Merkmale[3] von Organismen, die davon ausgehenden Kanten stellen die zugehörigen Merkmalsausprägungen dar. Jeder innere Knoten muss dabei mindestens zwei ausgehende Kanten besitzen. Die Blätter bezeichnen die so identifizierten Taxa.

Bei der Anwendung eines Bestimmungsschlüssels wird die Menge der infrage kommenden Taxa schrittweise reduziert. Dazu wird bei jedem Schritt ein Merkmal ausgewählt. Dann werden für die beobachteten Merkmalsausprägungen mit denen der Taxa verglichen und die diejenigen ausgeschlossen, für die sich dabei keine Schnittmenge ergibt.

Es existieren verschiedene Varianten von Bestimmungsschlüsseln.

1. single-access: Die Reihenfolge der abgefragten Merkmale ist fest vorgegeben (Abb. 3.5).
   a) diagnostic: Die Reihenfolge der Merkmale wird so gewählt, dass eine möglichst einfache und sichere Bestimmung erfolgen kann.
   b) synoptic: Die einzelnen Bestimmungschritte orientieren sich an der Systematik.
2. multi-access: Der Nutzer kann die Reihenfolge der Merkmale auswählen.
   a) tabular, matrix: Druckbare Darstellungsformen
   b) interactive: Computer-gestützt

2.1.3. Begriffsdefinitionen

Merkmal ist ein Konzept für das ein Beobachtungs- oder Messverfahren festgelegt wurde, welches wiederholbare Ergebnisse eines bestimmten Typs erzeugt [4].

Merkmalsausprägung ist das Ergebnis der Anwendung eines Beobachtungs- oder Messverfahrens für ein Merkmal [5].

[3] Englisch: character.
[4] Vgl. Hagedorn (2007) S.32.
[5] Vgl. Hagedorn (2007) S.33.

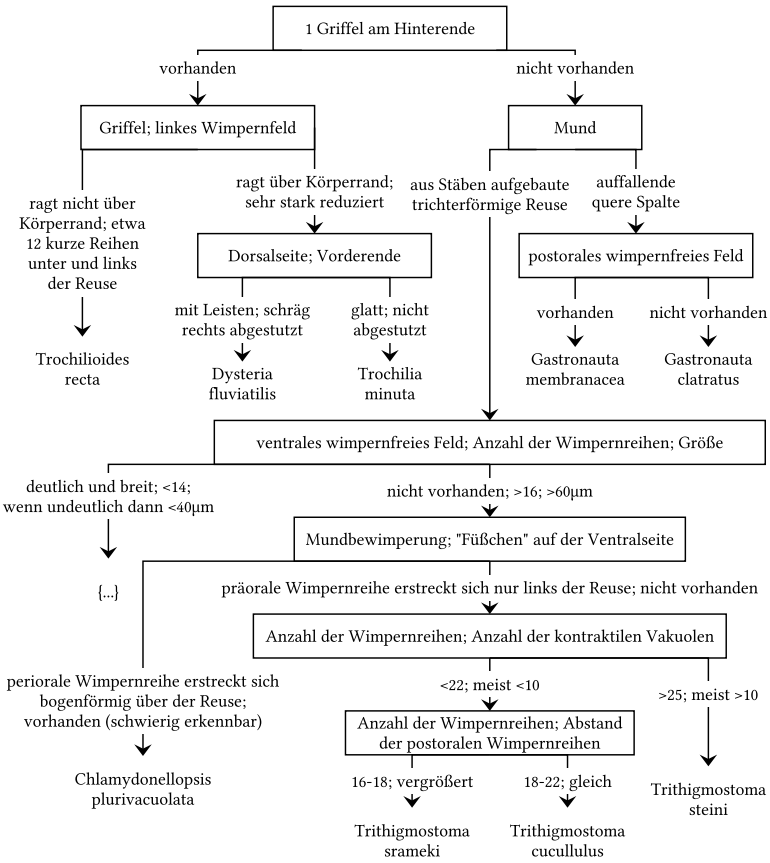

Abbildung 2.1.: Ausschnitt aus einem Bestimmungsschlüssel (Cyrtophorida), (vgl. Foissner u. a., 1991, S. 43)

2.2. Anwendungsspezifisches Wissen

Die Wissensbasis ist die Grundlage für die Bestimmung des Taxons eines beobachteten Mikroorganismuses. Sie enthält Inferenzregeln für das Schließen von Beobachtungen auf die für das Saprobiensystem relevanten und in der DIN-Norm festgelegten Taxa. Dabei muss sie insbesondere die Fachsprache der in der Norm referenzierten Quellen berücksichtigen. Die Quellen beschreiben einzelne Taxa in der Form, dass sie für geeignete Merkmale alle möglichen Merkmalsausprägungen aufzählen oder als Intervall angeben.

Beispiel: Differentialdiagnose Aspidisca cicada

1. Größe in vivo 25-40 x 20-40 µm.
2. Gestalt im Umriß rundlich-dreieckig, rechter Körperrand konvex, linker ziemlich gerade, häufig mit einem gerundeten Vorsprung in der Mundgegend, hinten quer abgestutzt. Wenig abgeflacht. Ventral eben, mit einem langen, hinten zugespitzten Pellicula-Dorn zwischen den 2 linken und den 3 rechten Transversalcirren [...]. Dorsal konvex, mit meist 6 - 8 längsverlaufenden Rippen. Anzahl und Höhe der Rippen stark variabel [...]. Pellicula starr.
3. Makronucleus hufeisenförmig, in vivo etwa 3-5 µm breit [...].
4. Kontraktile Vakuole und Exkretionsporus am rechten Körperrand etwa auf der Höhe der rechten Transversalcirren [...].
5. Anordnung und Zahl der Cirren konstant: 7 etwa 10 µm lange Ventralcirren in einer entlang dem Rücken verlaufenden Reihe von 4 und einer hinteren Reihe von 3 Cirren. 5 etwa 15 µm lange Transversalcirren, von denen die beiden linken rechts vom Mund in einer flachen Senke inserieren, die 3 rechten hintereinander entlang des ventralen Dornes [...]. 5 (selten 6) dorsale Wimpernreihen mit kurzen, borstenartigen Wimpern (nur nach Silberimprägnation gut erkennbar; [...]). Keine Caudalcirren.
6. Oralapparat zweigeteilt: vorne links, etwas überdeckt vom vordersten Ventralcirrus, 3 winzige, etwa 5 µm lange Frontalmembranellen in einer tiefen Grube; an der linken hinteren Ecke der Ventralfläche eine schräg zur Körperlängsachse gestellte, oval-schüsselförmige Mundhöhle mit meist 11 adoralen Membranellen [...].
7. ...

(vgl. Foissner u. a., 1991, S. 370)

3. Grundlagen

3.1. Wissensbasierte Systeme

> Wissensbasierte Systeme sind Programme, die auf der Grundlage von Wissen über einen bestimmten Anwendungsbereich Schlussfolgerungen ziehen können, und die so einem Benutzer helfen, ein Problem zu lösen oder eine Entscheidung zu treffen. (Borgelt u. a., 2003, S. 291)

Eine wesentliche Eigenschaft eines wissensbasierten Systems ist die Trennung in das Wissen über den Anwendungsbereich, das in *Wissensbasis* gespeichert ist, und die Problemlösungsmethoden zur Verarbeitung des Wissens, die in einer anwendungsunabhängigen Komponente enthalten sind[1]. Dadurch wird die Wiederverwendbarkeit der Problemlösungskomponente und die Möglichkeit zur Anwendung verschiedener Problemlösungsmethoden auf eine Wissensbasis erreicht.

3.1.1. Expertensysteme

Expertensysteme sind wissensbasierte Systeme, bei denen das Wissen von Experten stammt[2], im Gegensatz z. B. zu Wissen aus Messungen oder statistischen Erhebungen, und die die Schlussfolgerungsfähigkeit von Experten nachbilden soll. Expertensysteme existieren für verschiedene Aufgabenbereiche, wie z. B. für Analyse und Beratung im Finanzwesen oder der Geologie oder für Fehlerdiagnose. Ein System, welches nur die Problemlösungskomponente ohne Wissensbasis bereitstellt, wird als *Expertensystem-Shell* bezeichnet. Ein Beispiel für ein solches System ist d3web[3].

3.1.2. Decision Support Systeme

Entscheidungsunterstützende Systeme können sowohl Expertenwissen, als auch Wissen aus Falldaten, Messungen, etc. verwenden. Außerdem unterscheiden sie sich in ihrer Zielsetzung

[1] Beierle und Kern-Isberner (2003), Puppe (1990)
[2] Puppe (1990), S. 2.
[3] http://www.is.informatik.uni-wuerzburg.de/forschung/anwendungen/d3web/

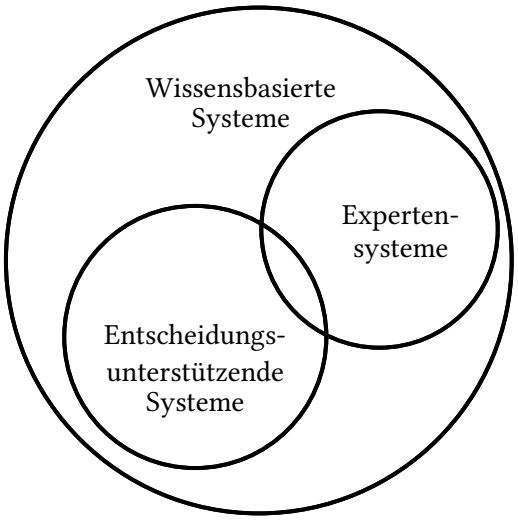

Abbildung 3.1.: Arten wissensbasierter Systeme

von Expertensystemen. Entscheidungsunterstützende Systeme sollen die Qualität der Entscheidungen des Anwenders verbessern, indem sie das Wissen sinnvoll aufbereiten und zur Verfügung stellen.

3.1.3. Wissen

Für den Begriff Wissen existieren sehr verschiedene Interpretationen. Oft wird Wissen als die Menge der als wahr betrachteten Aussagen definiert[4]. Im betrachteten Anwendungsfall ist das Wissen gegeben durch die Menge der Aussagen in der Fachliteratur, auf die in der DIN-Norm verwiesen wird.

Viele dieser Aussagen beinhalten Unsicherheiten, so dass Wissen nicht immer ausschließlich durch logische Aussagen und Regeln ausgedrückt werden kann, sondern diese Aussagen zusätzlich mit Werten zur Beschreibung der Unsicherheiten verknüpft werden müssen. Borgelt u. a. (2003) beschreiben drei Eigenschaften für unsicheres Wissen: *Impräzision*, *Unsicherheit* und *Vagheit*.

Impräzision. Wird Wissen als Menge von Alternativen angegeben, spricht man von Impräzision. Beispiele hierfür sind z. B. *„Gestalt kelch- oder verkehrt birnenförmig“* oder *„21-25 longitudinale Wimpernreihen“*. Impräzisionen lassen sich in logische Aussagen überführen. Handelt

[4]Vgl. Borgelt u. a. (2003), S.292.

es sich um eine endliche Menge von Alternativen, lässt sich diese durch einfache Disjunktion darstellen („*Gestalt = kelchförmig* $\vee$ *Gestalt = verkehrt birnenförmig*“). Intervalle können mit Hilfe von Vergleichsprädikaten beschrieben werden („*Anzahl der longitudinalen Wimpernreihen* $\geq 21 \wedge$ *Anzahl der longitudinalen Wimpernreihen* ≤ 25“).

Unsicherheit. Wenn in einer Aussage nicht alle möglichen Alternativen genannt werden können, also nicht genannte Alternativen nicht definitiv ausgeschlossen werden können, ist sie unsicher. Unbestimmte Numerale in den natürlichsprachlichen Aussagen, wie „meist“ oder „selten“, deuten auf Unsicherheiten. Ein Beispiel ist „*Makronucleus meist ellipsoid*“. Reicht die Kennzeichnung einer Aussage als sicher oder unsicher, kann sie in der *Modallogik* durch die *Modalitäten* „notwendig“ und „möglich“ beschrieben werden. Sollen Aussagen jedoch vergleichbar sein, ist ein Maß erforderlich. Es existieren verschiedene Ansätze, um den Grad der Sicherheit oder des Vertrauens zu quantifizieren, z. B. durch Wahrscheinlichkeiten oder Sicherheitsfaktoren.

Vagheit. Sind Alternativen in Aussagen nicht klar abgegrenzt, z. B. „*Oralapparat klein*“, bezeichnet man die Aussage als vage. Die Vagheit lässt sich durch die Definition klarer Grenzen beseitigen oder mithilfe der *Fuzzy-Logik* verarbeiten.

3.2. Unsicheres Schließen

Schlussfolgern bedeutet, aus bekanntem Wissen neues Wissen abzuleiten. Für eine derartige Ableitung wird als wahr bekanntes Wissen, die *Prämissen*, durch eine Inferenzrelation mit neuem Wissen, der *Konklusion*, verbunden.

Wenn eine Inferenzrelation durch eine Unsicherheit charakterisiert ist, also die damit abgeleiteten Konklusionen nicht stets war sind, ist es notwendig, die Richtigkeit der Schlussfolgerungen zu bewerten. In den Beschreibungtexten geschieht dies in der Regel sprachlich durch Begriffe wie z. B. „selten“, „oft“ oder „meist“. Wenn diese Ausdrücke durch einen Experten beziffert werden, können daraus Bewertungen für die Schlussfolgerungen berechnet werden. Dazu existieren verschiedene Verfahren.

3.2.1. Probabilistisches Schließen

Ein auf der mathematischen Wahrscheinlichkeitstheorie basierender Ansatz ist das *Probabilistische Schließen*. Bei diesem Ansatz wird die Wahrscheinlichkeit, dass eine Hypothese H unter einer Bedingung B zutrifft, mithilfe des Bayes'schen Theorems berechnet.

Zur Überprüfung der Anwendbarkeit des probabilistischen Ansatzes wird das Beispiel „Die kontrahierten Zooide der Art *Epistylis entzii* sind meist ellipsoid“ untersucht. Diese Aussage lässt sich als Implikation $T_{E.entzii} \xrightarrow{p_{meist}} M_{ellipsoid}$ der Aussagen

$$T_{E.entzii} = \text{„Ein Mikroorganismus } \omega \text{ gehört zur Spezies E. entzii“ und}$$
$$M_{ellipsoid} = \text{„Die Gestalt des Mikroorganismus } \omega \text{ ist ellipsoid“}$$

mit der bedingten Wahrscheinlichkeit $P(M_{ellipsoid}|T_{E.entzii}) = p_{meist}$ ausdrücken[5]. Da bei der Bestimmung von Mikroorganismen aber von beobachteten Merkmalen auf die Spezies geschlossen werden muss, ist $P(T_{E.entzii}|M_{ellipsoid})$ die gesuchte Wahrscheinlichkeit. Diese ließe sich durch

$$P(T_{E.entzii}|M_{ellipsoid}) = \frac{P(M_{ellipsoid}|T_{E.entzii})P(T_{E.entzii})}{P(M_{ellipsoid})} = \frac{p_{meist}P(T_{E.entzii})}{P(M_{ellipsoid})}$$

berechnen.

Im Kontext des Verfahrens zur Berechnung des Saprobienindex sind die Häufigkeiten der Spezies jedoch gesuchte Größen, die gerade durch die Bestimmung beobachteter Organismen ermittelt werden sollen. $P(T_{E.entzii})$ ist also unbekannt. Zudem kann angenommen werden, dass $P(M_{ellipsoid})$ nicht erhoben wird und somit auch nicht zur Verfügung steht. Die für den Anwendungsfall gegebene Wissensbasis lässt also kein probabilistisches Schließen zu.

3.2.2. Sicherheitsfaktoren

Das Konzept der *Sicherheitsfaktoren* beruht im Gegensatz zum probabilistischen Schließen nicht auf tatsächlichen Wahrscheinlichkeiten, sondern auf Vertrauensgraden, die eine subjektive Wahrscheinlichkeit der Gültigkeit einer Hypothese $P(H)$ beschreiben. Dazu werden Sicherheitsfaktoren $CF(H|E) \in [-1, 1]$ angegeben, die als Maß für die Änderung dieser subjektiven Wahrscheinlichkeit $P(H|E) - P(H)$ verstanden werden können. Hier wird deutlich, dass der Experte für die Angabe eines solchen Sicherheitsfaktors eine Vorstellung von der bedingten Wahrscheinlichkeit $P(H|E)$ haben muss.

Bezogen auf die Frage nach dem Taxon eines Mikroorganismus heißt das aber, dass ein Maß dafür angegeben werden muss, wie sich die Wahrscheinlichkeit für ein Taxon unter der Voraussetzung eines bestimmten Merkmals ändert. Die Fachtexte liefern aber bestenfalls Hinweise über die Wahrscheinlichkeit eines Merkmals für ein Taxon, z. B. durch eine Angabe wie

[5] Ein Experte könnte p_{meist} z. B. auf 0,9 schätzen.

„Farbe meist grün, selten blau“. Aufgrund dieser umgekehrten Schlussrichtung eignen sich Sicherheitsfaktoren nicht für den speziellen Anwendungsfall.

3.2.3. Dempster-Shafer-Theorie

Atkinson und Gammerman (1987) verwenden die Funktionen der *Dempster-Shafer-Theorie* für die Bewertung von Hypothesen bei der biologischen Bestimmung. Die Dempster-Shafer-Theorie basiert auf einem *Basismaß*

$$m : \mathcal{P}(\Omega) \rightarrow [0, 1],$$

wobei Ω eine Menge von möglichen, sich gegenseitig ausschließenden Elementarereignissen ist, die auch als *Rahmen* bezeichnet wird. Die betrachteten Ereignisse lassen sich als Untermengen von Ω darstellen, also auch als Elemente der Potenzmenge $\mathcal{P}(\Omega)$. $m(A)$ gibt dann das Maß an Glauben an, dass exakt der Menge A zugewiesen wird.

Die Funktion für das Basismaß muss den Bedingungen

$$m(\emptyset) = 0 \tag{3.1}$$

$$\sum_{A \subseteq \Omega} m(A) = 1 \tag{3.2}$$

genügen.

Des Weiteren wird eine Glaubensfunktion

$$Bel(A) = \sum_{B \subseteq A} m(B)$$

und eine Plausibilitätsfunktion

$$Pl(A) = 1 - Bel(\overline{A}) = \sum_{A \cap B \neq \emptyset} m(B)$$

definiert. Die Glaubensfunktion bestimmt dabei, inwieweit Evidenzen eine Hypothese unterstützen und die Plausibilitätsfunktion zeigt an, in welchem Maße die Evidenzen der Hypothese nicht widersprechen. Für die Kombination von Basismaßen, die sich aus unterschiedlichen Evidenzen ergeben, existieren verschiedene Regeln.

Praktisch heißt das für die Bestimmung von Mikroorganismen jedoch, dass Evidenzwerte notwendig sind, die angeben, in welchem Maße einer Regel „Wenn ein bestimmtes Merkmal vorliegt, ist die Lösung Element einer genau definierten Untermenge der Menge aller mögli-

chen Lösungen“ vertraut wird. Derartige Evidenzen sind den Beschreibungen jedoch nicht zu entnehmen.

3.2.4. Schwellenwert für Evidenzen

Aus Evidenzen, wie sie sich aus der Wissensgrundlage ergeben, lassen sich also keine vergleichbaren Bewertungen der Lösungen ableiten. Wird eine Merkmalsausprägung (oder eine Menge alternativer Merkmalsausprägungen) mit nicht sicherer Evidenz als sicher betrachtet, können Lösungen fälschlicherweise ausgeschlossen werden, da offenbar nicht alle Alternativen bekannt sind. Um solche falsch negativen Schlüsse zu verhindern, müssen Merkmale eines Taxons, deren Evidenz unsicher ist, als unbekannt angesehen werden.

Verwirft man unsichere Merkmale auf diese Art jedoch ganz, wird u. U. eine Möglichkeit vergeben, Organismen voneinander zu unterscheiden. Es kann daher sinnvoll sein, Evidenzen ab einem bestimmten Schwellenwert als sicher zu betrachten, wenn die dadurch entstehende Wahrscheinlichkeit von falsch negativen Ergebnissen vertretbar gering ist.

Unter der Annahme, dass das mikroskopische Bild in der Regel mehrere Exemplare eines Taxons zeigt und der Anwender diese Zugehörigkeit zum selben Taxon auch erkennt, erhöht sich die Wahrscheinlichkeit, dass ein Merkmal innerhalb dieser Gruppe erkannt wird. Dies wird im Rahmen der Saprobienindexberechnung insbesondere durch die erforderliche Angabe einer Abbundanzziffer offenbar, die eine Stufe der Individuendichte angibt, die bis zu mehrere hundert Exemplare innerhalb einer Probe betragen kann.

Nimmt man weiterhin an, dass der Anwender ein Merkmal, das mit der Wahrscheinlichkeit $p = P(M|T)$ bei Exemplaren eines Taxons auftritt, erkennt, wenn es bei mindestens m von n beobachteten Exemplaren dieses Taxons ausgeprägt ist, lässt sich die Fehlerwahrscheinlichkeit durch

$$P(\text{„}falsch\ negativ\text{“}) = \sum_{k=0}^{k<m} \binom{n}{k} p^k (1-p)^{n-k}$$

mit

n: Anzahl der gleichzeitig beobachteten Exemplare eines Taxons

m: Anzahl der Exemplare, bei denen das Merkmal für eine Erkennung ausgeprägt sein muss

k: Anzahl der Exemplare, bei denen das gesuchte Merkmal ausgeprägt ist

p: die Wahrscheinlichkeit, dass das Merkmal bei einem Exemplar des Taxons ausgeprägt ist

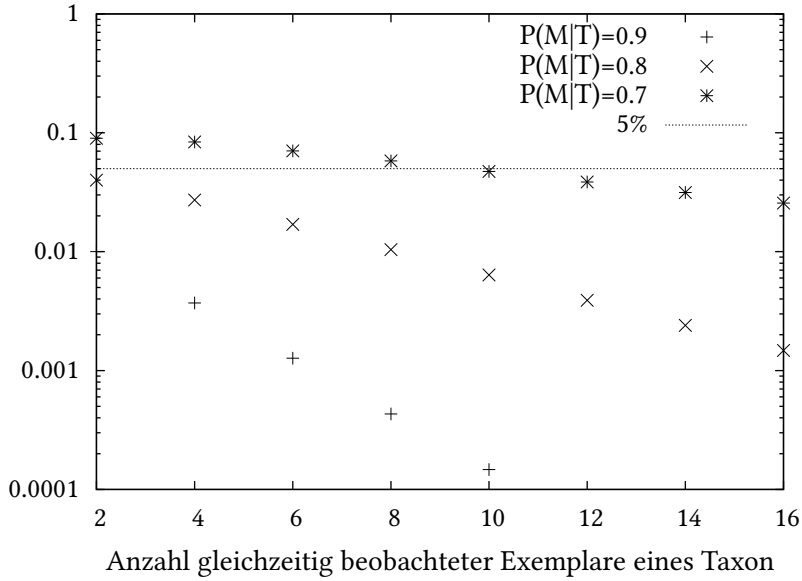

Abbildung 3.2.: Wahrscheinlichkeit für falsch negative Ergebnisse für $m = ceil(\frac{n}{2})$

approximieren[6]. Abbildung 3.2 zeigt die Fehlerwahrscheinlichkeiten für den Fall, dass das Merkmal bei mindestens der Hälfte der Exemplare ($m = ceil(\frac{n}{2})$) ausgeprägt sein muss, damit es erkannt wird.

Daraus wird ersichtlich, dass in der Praxis ein Schwellenwert von $\approx 0,9$, bei sehr häufigen Organismen (Kolonien) auch kleiner, eine hinreichend sichere Bestimmung ermöglicht.

3.3. Problemklasse Klassifikation

Um bekannte Konzepte zur Lösung von Problemen in wissensbasierten Systemen nutzen zu können, ist zunächst die Problemklasse zu identifizieren. Puppe (1990) unterteilt die Problemtypen in die Klassen *Klassifikation*, *Konstruktion* und *Simulation* und definiert die Problemklasse Klassifikation über folgende Eigenschaften:

1. Der Problembereich besteht aus zwei endlichen, disjunkten Mengen von Problemmerkmalen und Problemlösungen und aus typischerweise unsicherem,

[6]Es handelt sich um eine hypergeometrische Verteilung, wobei alle existierenden Exemplare des Taxons die Grundgesamtheit bilden und die gleichzeitig beobachteten Exemplare des Taxons die Stichprobe darstellen. Daher ist $N \gg n$ und die Approximation durch die Binomialverteilung ist möglich.

mehrstufigem Wissen über die Beziehungen zwischen Problemmerkmalen und Problemlösungen.

2. Ein Problem ist durch eine eventuell unvollständig gegebene Teilmenge von Problemmerkmalen charakterisiert.
3. Das Ergebnis der Klassifikation ist die Auswahl einer oder mehrerer der Problemlösungen.
4. Wenn die Qualität der Problemlösung durch Erfassung zusätzlicher Problemmerkmale verbessert werden kann, so ist es eine Teilaufgabe der Klassifikation, zu bestimmen, ob und welche zusätzlichen Problemmerkmale angefordert werden sollen.

(Puppe, 1990, S. 42)

Bei der Bestimmung von Mikroorganismen stellen die beobachteten Merkmalsausprägungen eines Individuums die Menge von Problemmerkmalen dar. Die bekannten Taxa sind die Menge der Problemlösungen. Das Wissen über die Beziehungen zwischen Problemmerkmalen und -lösungen ergibt sich aus der Zuordnung bestimmter, möglicher Merkmalsausprägungen zu den Taxa, die den Beschreibungen in der Fachliteratur entnommen werden. Für die Durchführung der Bestimmung sollen möglichst wenige Merkmale des Organismus abgefragt werden. Als Ergebnis der Bestimmung sollen alle infrage kommenden Taxa ausgegeben werden. Damit hat die Aufgabenstellung alle Eigenschaften der Klassifikation.

Aus Anwendersicht handelt es sich um eine Zuordnung eines physikalischen Objektes zu einer Objektklasse aufgrund seiner beobachteten Merkmale. Derartige Probleme werden unter dem Problemtyp *Objekt-Identifikation* zusammengefasst[7].

3.4. Evaluierung der Problemlösungsmethoden

Auf Basis der Aussagen- oder Prädikatenlogik werden 4 Basis-Problemlösungsmethoden für die Klassifikation benannt. Die Methoden verwenden Regeln, also formalisierte Konditionalsätze der Form „*Wenn* $Prämisse$ *dann* $Konklusion$“, um Lösungen zu generieren und zu bewerten.

Vorwärtsverkettung: Die Vorwärtsverkettung wertet die Regeln ausgehend von den eingegebenen Problemmerkmalen aus und schließt so auf die Lösungen.

[7] Vgl. Puppe (1990), S. 47.

Rückwärtsverkettung: Die Rückwärtsverkettung bewertet eine Lösung durch Auswertung aller Regeln, die zu dieser Lösung führen.

Establish-Refine: Die Establish-Refine-Methode setzt hierarchische Lösungsklassen voraus. Lösungsklassen werden bei dieser Methode rekursiv durch Rückwärtsverkettung bestätigt und so verfeinert.

Hypothesize-and-Test: Bei der Hypothesize-and-Test-Methode werden zunächst Lösungskandidaten (Hypothese) durch Vorwärtsverkettung generiert, die dann durch Rückwärtsverkettung überprüft werden (Test).

Basis-Problemlösungsmethoden sind für viele Arten der Wissensrepräsentation einsetzbar, aber nicht optimal in Bezug auf Wissenserwerb und Effizienz. Teilweise basierend auf den Basis-Problemlösungsmethoden existieren abhängig von der Wissensart präzisere Methoden, die auch als starke Problemlösungsmethoden bezeichnet werden.

3.4.1. Überblick

Unterschieden werden die Wissensarten: Fälle, Erfahrungswissen und Modelle. Fälle sind dabei Beobachtungen früherer Probleme und ihre Zuordnungen zu Lösungen. Erfahrungswissen beinhaltet die Abbildung mentaler Modelle von Experten. Wichtige starke Problemlösungsmethoden der Klassifikation lassen sich nach den Wissensarten wie folgt unterteilen[8]:

Erfahrungswissen Sichere oder Heuristische Klassifikation

- sichere Klassifikation
- heuristische Klassifikation

Modelle Kausale Klassifikation

- überdeckende Klassifikation
- funktionale Klassifikation

Fälle Fälle-orientiertes Schließen

- statistische Klassifikation
- fallbasiertes Schließen
- Neuronale Netze

[8] Vgl. Puppe u. a. (2003), S. 620 und Puppe (1990), S. 52.

Diese Methoden haben unterschiedliche Eigenschaften und Anforderungen an die Wissensbasis und werden im Folgenden in Bezug auf die Anwendungsdomäne evaluiert, da nur Methoden eingesetzt werden können, deren Voraussetzungen durch die Wissensbasis erfüllt werden. Das Wissen, dass in Beschreibungen von Taxa zur Verfügung steht, hat immer die Form *Wenn <Taxon>, dann <Merkmalsausprägung>*. Dabei stellt jedes Taxon eine mögliche Lösung dar. Dieses Wissen ist einerseits oft impräzise, es können also mehrere alternative Ausprägungen für ein Merkmal existieren, andererseits gelegentlich aber auch unsicher oder vage.

3.4.2. Sichere Klassifikation

Die Auswertung von sicherem Wissen und sicheren Daten der Form: Merkmal bewirkt Lösung ($M \Rightarrow L$) wird als sichere Klassifikation bezeichnet. Derartige Wissensbasen lassen sich z. B. durch Entscheidungsbäume oder -tabellen darstellen.

Die Regeln der Form *Wenn <Lösung>, dann <Merkmal>* lassen sich für die sichere Klassifikation nicht direkt anwenden. Allerdings lassen sich mit der Inferenzregel des *Modus tollens* $((L \Rightarrow M) \wedge \neg M) \Rightarrow \neg L$ Verdachtshypothesen ausschließen. Impräzises Wissen kann für die sichere Klassifikation durch Disjunktion von Regeln verwendet werden, allerdings bietet sie keine Möglichkeit, Unsicherheiten oder Vagheiten zu berücksichtigen.

3.4.3. Heuristische Klassifikation

Im Gegensatz zur sicheren Klassifikation benutzt die heuristische Klassifikation unsicheres Wissen der Form: Merkmal bewirkt Lösung mit Unsicherheit x. Die Beziehung zwischen Merkmal und Lösung wird dabei zusätzlich durch einen *positiven* und einen *negativen Vorhersagewert* charakterisiert. Der positive Vorhersagewert gibt an, wie stark die Beobachtung eines Merkmals die Lösung unterstützt ($P(L|M)$) und der negative, wie stark die Abwesenheit gegen die Lösung spricht ($P(\neg L|\neg M)$). Diese Unsicherheiten werden von Experten geschätzt. Für die Verknüpfung dieser Werte existieren verschiedene Ansätze.

Wie bei der sicheren Klassifikation können die Regeln der Wissensbasis nicht direkt angewendet, aber Lösungkandidaten durch Modus tollens ausgeschlossen oder mit negativem Vorhersagewert bewertet werden. Impräzisionen werden bei dieser Art der Klassifikation ebenso berücksichtigt wie Unsicherheiten, Vagheiten jedoch nicht.

3.4.4. Überdeckende Klassifikation

Sind die Lösungen zuverlässig durch charakteristische Mengen von Problemmerkmalen beschreibbar, eignet sich die Methode der überdeckenden Klassifikation. Die Wissensbasis enthält Merkmale, Lösungen und kausale Regeln der Form: Lösung verursacht Merkmal ($L \Rightarrow M$), wobei bestimmte Merkmale Zwischenzustände darstellen können, die ihrerseits weitere Merkmale bedingen.

Die Lösungen können dabei also nicht direkt aus den beobachteten Merkmalen abgeleitet werden, sondern es müssen Verdachtshypothesen überprüft werden, z. B. indem alle möglichen Lösungen verdächtigt werden. Die Bewertung einer Lösung ist umso besser, je mehr beobachtete und je weniger nicht-beobachtete Merkmale sie herleitet. Dabei können Unsicherheiten in Form von Evidenzwerten berücksichtigt werden.

3.4.5. Funktionale Klassifikation

Bei der funktionalen Klassifikation wird ein System untersucht, dessen Komponenten und deren Verhalten anhand eines funktionalen Modells bewertet wird[9]. Sie kommt z. B. bei der Fehlersuche zum Einsatz.

3.4.6. Statistische Klassifikation

Die statistische Klassifikation lässt sich auf große, repräsentative Sammlungen von Fällen anwenden, denen eine Lösung zugeordnet wurde. Sie unterscheidet sich von der heuristischen Klassifikation dadurch, dass die Unsicherheiten in den Beziehungen zwischen Problemmerkmalen und Lösungen nicht von Experten geschätzt, sondern statistisch erhoben werden. Die Bewertung von Lösungen wird mithilfe des Bayes-Theorems berechnet. Dazu müssen die a priori-Wahrscheinlichkeiten der Lösungen bekannt sein, sowie die bedingten Wahrscheinlichkeiten für das Auftreten eines Merkmals unter der Voraussetzungen einer bestimmten Lösung.

$$P(L_i|M_1 \wedge ... \wedge M_m) = \frac{P(L_i)P(M_1|L_i)...P(M_m|L_i)}{\sum_{j=1}^{n} P(L_j)P(M_1|L_j)...(P(M_m|L_j} \tag{3.3}$$

Weitere Voraussetzungen dafür sind, neben der Repräsentativität der Fallsummlung, eine vollständige und disjunkte Lösungsmenge und die Unabhängigkeit der Merkmale.

[9] Puppe (1990)

3.4.7. Fallbasiertes Schließen

Fallbasiertes Schließen verwendet ein Ähnlichkeitsmaß, um Fälle aus der Falldatenbank mit Beobachtungen zu vergleichen. Die Lösungen der Vergleichsfälle werden anhand der Ähnlichkeit bewertet und gegebenenfalls übernommen.

3.4.8. Neuronale Netze

Neuronale Netze verwenden Fallsammlungen für die Adaption ihrer Gewichte und setzen dabei verschiedene Lernregeln und Netztopologien ein[10]. Sie eignen sich vor allem für Anwendungsbereiche, in denen kein explizites Wissen über das zu lösende Problem vorliegt.

3.4.9. Fazit

Das für die Bestimmung von Mikroorganismen zur Verfügung stehende Wissen aus der Fachliteratur ist zwar aus Fällen entstanden, nämlich aus Proben von Individuen, deren Merkmale von Biologen untersucht wurden und die dann einem bestimmten Taxon zugeordnet wurden. Im Zuge dessen wurde es aber aggregiert und in eine Beschreibung überführt, die den Anspruch erhebt, für alle Mitglieder eines Taxons gültig zu sein. Dazu werden Impräzisionen und Unsicherheiten eingeführt, die eine Rekonstruktion der zugrundeliegenden Fälle unmöglich machen, so dass eine Klassifikation auf der Basis von Fällen nicht möglich ist. Vor allem aber die Tatsache, dass die Häufigkeiten der Taxa und somit die a priori-Wahrscheinlichkeiten der Lösungen nicht bekannt sind, lässt die statistische Klassifikation als Lösungsmethode ausscheiden.

Die Methode der funktionalen Klassifikation ist offensichtlich ebenfalls nicht einsetzbar, da die Menge aller Mikroorganismen nicht als ein System angesehen werden kann, für das irgendeine Art von Modell vorliegt.

Wenn Unsicherheiten vernachlässigt werden können ist es möglich, die sichere Klassifikationsmethode einzusetzen. Sind sinnvolle Evidenzen bekannt, eignet sich auch die Methode der heuristischen Klassifikation. Die überdeckende Methode ist in jedem Falle einsetzbar.

3.5. Symbolic Data Analysis

Unter dem Begriff *Symbolic Data Analysis (SDA)* werden Methoden zusammengefasst, die der Erzeugung, Beschreibung und Verarbeitung komplexer Daten dienen. Ziel der SDA ist die Ge-

[10] Vgl. Puppe u. a. (2003) S.621.

neralisierung von Verfahren des Data Mining und statistischer Methoden für die Verwendung mit Einheiten höherer Ebene, die durch sogenannte symbolische Daten beschrieben werden[11]. Solche Verfahren sind dann sinnvoll, wenn die Daten bereits in aggregierter Form vorliegen, wie im Fall der Beschreibung von Mikroorganismen. Aber auch die schiere Menge von Daten kann die Anwendung klassischer Methoden der Datenanalyse erschweren und eine Zusammenfassung notwendig machen. Zudem kann auch der Datenschutz eine Zusammenfassung individueller Daten erfordern.

Eine derartige Aggregation von Daten über eine bestimmte Gruppe statistischer Einheiten ist bei der klassischen Datenanalyse, z. B. mit Mitteln der deskriptiven Statistik, in der Regel mit einem Verlust von Informationen über die Variabilität der Daten verbunden. Bei der SDA sollen diese Informationen in den aggregierten Daten erhalten bleiben und verarbeitet werden können.

3.5.1. Prinzipien der SDA

Diday (2008) benennt u. a. folgende wesentliche Prinzipien der SDA[12]:

1. Die SDA beruht auf zwei Ebenen von Einheiten. Die erste Ebene beinhaltet statistische Einheiten, wie z. B. Individuen, die als *"first level units"*[13] bezeichnet werden. *„Second level units"* beschreiben *Konzepte*[14].

2. Ein Konzept wird gebildet durch eine symbolische Beschreibung, z. B. der charakteristischen Merkmale, eine Klasse von Objekten *(Intension)* oder durch eine Menge von Individuen *(Extension)*[15], die durch Aggregation der Beschreibungen der Individuen erzeugt wird. Dieser Aggregationsprozess wird auch als *Generalisierung* bezeichnet.

3. Die Beschreibung muss die Variationen der Objekte der Klasse berücksichtigen, z. B. durch die Abbildung von Mengen, Intervallen, Verteilungen usw.

4. Konzepte werden durch sogenannte symbolische Objekte modelliert. Symbolische Objekte können einem Lernprozess unterliegen.

5. Die SDA erweitert die explorative Datenanalyse und die Verfahren des Data Mining für den Fall, dass die statistischen Einheiten durch symbolische Daten beschrieben werden.

[11] Vgl. Diday (2008), S. 4.
[12] Vgl. Diday (2008) S.22.
[13] Gebräuchlich sind auch die Bezeichnungen *first-order* und *second-order*.
[14] Diday (2008), S. 20 verweist auf Aristoteles' Begriffe der „Ersten Substanz" (Beispiel: ein Pferd) und „Zweiten Substanz" (Beispiel: die Spezies Pferd).
[15] Vgl. Diday (2008) S.8.

6. Einige Verfahren der SDA erzeugen Ausgaben in Form neuer symbolischer Objekte.

3.5.2. Symbolische Variablentypen

Sei Ω eine Grundgesamtheit statistischer Einheiten und Δ der Beschreibungsraum der Elemente aus Ω. Δ basiert auf einer Menge von p Variablen (Eigenschaften, Merkmalen) $Y_1, ..., Y_p$, die jeweils eine Abbildung $Y_j : \Omega \rightarrow \Delta_j$ definieren, wobei Δ_j den Beschreibungsraum von Y_j darstellt. Dann kann die Beschreibung eines Individuums $\omega \in \Omega$ definiert werden als

$$\boldsymbol{Y}(\omega) = (Y_i(\omega), ..., Y_p(\omega)) \in \Delta = \Delta_1 \times ... \times \Delta_p$$ [16]

Sei O_j der Wertebereich einer Variable Y_j, so wird in der klassischen Datenanalyse die Variable Y_j für ein einzelnes Element aus Ω durch einen Wert aus O_j beschrieben. Es gilt also $\Delta_j \equiv O_j$ und die Beschreibung $\boldsymbol{Y}(\omega)$ lässt sich somit durch $\Delta = O_1 \times ... \times O_p$ ausdrücken.

Beispielsweise sei Ω die Menge aller Studenten mit den Eigenschaften Größe und Geschlecht mit $O_{\text{Größe}} = \mathbb{R}^+$ und $O_{\text{Geschlecht}} = \{\text{m}, \text{w}\}$. Dann werden einzelne Elemente aus Ω z. B. beschrieben durch $(1.82, \text{m})$ oder $(1.95, \text{w})$.

Um verschiedene Einheiten, wie z. B. Klassen von Objekten oder Expertenwissen, aber auch Unsicherheiten zu beschreiben, werden bei der SDA komplexere Beschreibungsräume Δ_j betrachtet. Zu diesem Zweck werden sogenannte symbolische Variablen eingeführt, wobei klassische quantitative oder qualitative Variablen Sonderfälle der symbolischen Typen sind (Abb. 3.3).

Auch symbolische Variablentypen werden in qualitative und quantitative unterteilt. Quantitative Variablen können einwertig sein, wenn sie genau einen Wert aus dem Wertebereich annehmen, mehrwertig, wenn sie durch eine endliche Untermenge dargestellt werden oder durch ein Intervall über dem Wertebereich ausgedrückt werden. Weiterhin können quantitative symbolische Variablen als Histogramme oder Funktionen dargestellt werden.

Qualitative Variablen können analog zu quantitativen ebenfalls ein- oder mehrwertig sein. Ist die Variable zudem ordinal, ist auch eine Intervalldarstellung möglich. Außerdem wird ein *qualitativ modaler Variablentyp* definiert, der jedem Wert des Wertebereichs eine absolute Häufigkeit oder Wahrscheinlichkeit zuordnet.

Eine einwertige symbolische Variable entspricht der Standard-Variable mit $\Delta_j \equiv O_j$, wie im obigen Beispiel. Mehrwertige symbolische Variablen haben dagegen einen Wert, der eine

[16] Vgl. Ciampi u. a. (2000), S. 788.

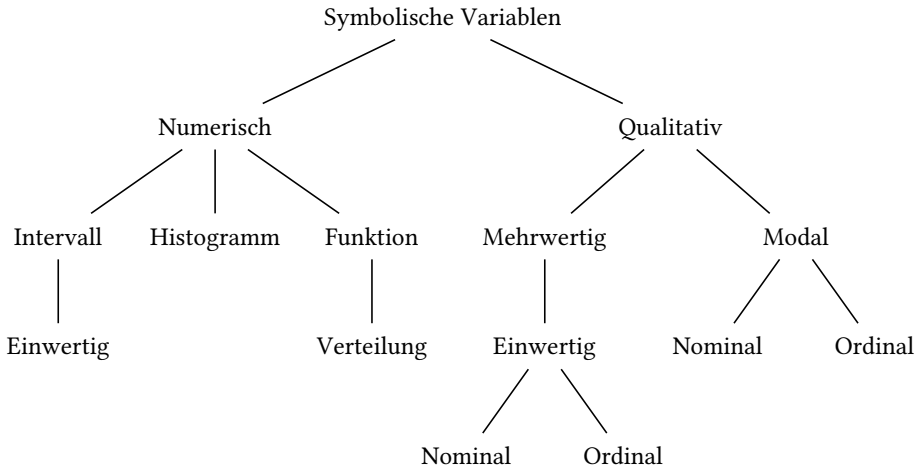

Abbildung 3.3.: Ontologie symbolischer Variablen (vgl. Noirhomme-Fraiture und Brito, 2011, S. 160)

Untermenge von O_j darstellt: $Y_j(\omega) = Q$ mit $Q \subseteq O_j$. Somit wird die Variable durch eine Abbildung $Y_j : \Omega \to \{Q \mid Q \subseteq O_j\}$ definiert und der Beschreibungsraum entspricht der Potenzmenge $\Delta_j = \{Q \mid Q \subseteq O_j\} = \mathcal{P}(O_j)$.

Existiert für den Wertebereich O_j eine Ordnungsrelation, also bei quantitativen oder ordinalen qualitativen Variablen, lassen sich Untermengen von O_j als Intervall $[l, u]$ mit $l, u \in O_j$ und $l \leq u$ darstellen. Die Menge aller möglichen Intervalle über O_j bildet dann den Beschreibungsraum $\Delta_j = \{[l, u] \mid l, u \in O_j \wedge l \leq u\} = [O_j]$.

3.5.3. Symbolische Objekte

Ein *symbolisches Objekt* S_C dient der Modellierung eines Konzeptes C und wird durch ein Tripel $S_C = (a_C, R_c, d_C)$ definiert[17]. Dabei ist d_C eine symbolische Beschreibung $d_C = (d_1, \ldots, d_p)$ aus dem Beschreibungsraum $D = D_1 \times \ldots \times D_p$ des Konzeptes[18] (Abb. 3.4).

a_C stellt eine Abbildung $a : \Omega \to L$ dar, die die Zugehörigkeit eines Objektes ω zu der durch das Konzept beschriebenen Klasse ausdrückt, wobei L z. B. deterministisch $L = \{wahr, falsch\}$ oder unsicher $L = [0, 1]$ sein kann. Dazu wird unter Verwendung der Relation R_C die Übereinstimmung von $\boldsymbol{Y}(\omega)$ mit d_C als $a_C(\omega) = [\boldsymbol{Y}(\omega)\ R_C\ d_C]$ definiert (Abb. 3.4). R_C besteht aus

[17] Vgl. Diday (2008) S. 25.
[18] Nicht zu verwechseln mit Δ_j, dem Beschreibungsraum für Y_j der „realen“ Objekte.

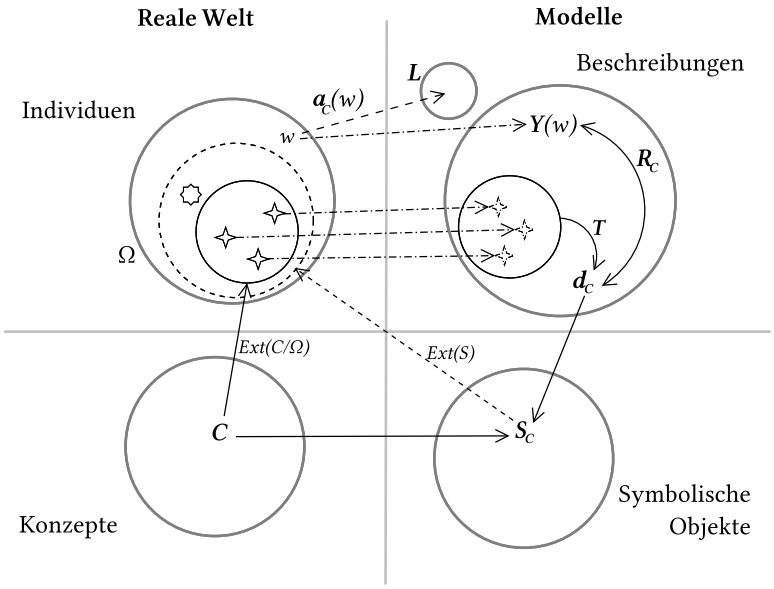

Abbildung 3.4.: Beziehungen zwischen Konzepten, Beschreibungen und symbolischen Objekten (vgl. Diday, 2008, S. 24)

einer Verknüpfung von Relationen R_j für die einzelnen Variablen Y_j. Die Relation R_j muss dabei entsprechend des bei der Generalisierung angewendeten Operators T gewählt werden. Sinnvolle Relationen für einzelne Variablen sind z. B. $R_j = \subseteq$ für $T = \cup$, $R_j = \supseteq$ für $T = \cap$.

Als Beispiel sei das Konzept „Student“ betrachtet. Durch Generalisierung der Daten der zu dieser Klasse gehörenden Individuen lässt sich eine als Intervall ausgedrückte Beschreibung der Eigenschaft „Größe“ in der Form $d_{\text{Größe}} = [min\{\boldsymbol{Y}_{\text{Größe}}(\omega) \mid \omega \in \Omega\}, max\{\boldsymbol{Y}_{\text{Größe}}(\omega) \mid \omega \in \Omega\}]$ erzeugen. Die Relation $R = \in$ zeigt dann an, ob die Beschreibung der Größe eines unbekannten Individuums x mit dem Konzept übereinstimmt: $a(x) = \boldsymbol{Y}(x) \in d_{\text{Größe}}$.

Event. Ein *Event* $e = [Y_j\ R_j\ d_j]$ ist die grundlegende Art eines symbolischen Objektes. Es definiert ein Objekt, dessen Beschreibung für die Variable Y_j durch d_j gegeben ist[19]. Das Konzept „männliche Studenten“ lässt sich z. B. durch ein Event $e = [Y_{\text{Geschlecht}} = d_{\text{Geschlecht}}]$ mit $d_{\text{Geschlecht}} = m$ und die Abbildung $e(\omega) = (Y_{\text{Geschlecht}}(\omega) \equiv m)$ darstellen.

Assertion. Ein symbolisches Objekt kann auch durch eine Konjunktion von n Events $S = e_1 \wedge \ldots \wedge e_n$ beschrieben werden. Derartige symbolische Objekte heißen *Assertion*.

[19] Vgl. Ciampi u. a. (2000) S. 790.

Intension und Extension. Die *Intension* eines symbolischen Objektes besteht aus der Menge aller Objekte des Konzeptes $Ext(C/\Omega)$ und wird durch seine Beschreibung d_C modelliert. Unter der *Extension* eines symbolischen Objektes versteht man die Menge aller Objekte, die durch das symbolische Objekt modelliert werden $Ext(S_C)$, also seiner Beschreibung entsprechen.

Ziel der Modellierung als symbolisches Objekt ist die Übereinstimmung von $Ext(C/\Omega)$ und $Ext(S_C)$. Inwieweit das der Fall ist, hängt offensichtlich von der Beschreibung d_C ab. Dieser Umstand ist auch an den vorgenannten Beispielen nachvollziehbar. Wird das Konzept „männliche Studenten“ nur durch die Variable „Geschlecht“ beschrieben, gilt augenscheinlich $Ext(C/\Omega) \subseteq Ext(S_C)$, da $Ext(S_C)$ alle männlichen Lebewesen aus Ω enthält und nicht nur die Studenten. Auch Abb. 3.4 zeigt ein derartiges Beispiel.

3.6. SDD - Structured Descriptive Data

Die Organisation *Biodiversity Information Standards*[20] definiert den Standard *Structured Descriptive Data (SDD)*[21] als XML-basiertes Datenmodell für die Beschreibung von Taxa. Im Folgenden werden wesentliche Komponenten des SDD-Datenmodells beschrieben.

Dataset. In den Datasets werden Informationen über die Taxa (TaxonNameSet), die verwendeten Merkmale (CharacterSet) und die Ausprägungen der Merkmale für die Taxa (CodedDescriptionSet) zusammengefasst. Zudem können sie u.a. Proben (SpecimenSet), vorgefertigte Bestimmungsschlüssel (IdentificationKeySet) und Medien (MediaObjectSet) enthalten.

3.6.1. Taxa

TaxonNameSet. Die Taxa werden im Datentyp TaxonNameSet verwaltet, der eine Liste von TaxonName-Elementen enthält. TaxonName-Elemente verfügen über ein *id*-Attribut, über das sie referenziert werden können, z. B. von den zugehörigen Beschreibungen. In Bezug auf die symbolische Datenanalyse bilden Taxa durch diese Verknüpfung die Konzepte.

3.6.2. Merkmale

CharacterSet. Die verwendeten Merkmale werden im *CharacterSet* aufgelistet. Das SDD-Schema definiert dazu die vier Merkmalstypen *CategoricalCharacter*, *QuantitativeCharacter*,

[20] Auch bekannt als Taxonomic Databases Working Group (TDWG) http://www.twdg.org.
[21] http://www.tdwg.org/standards/116/.

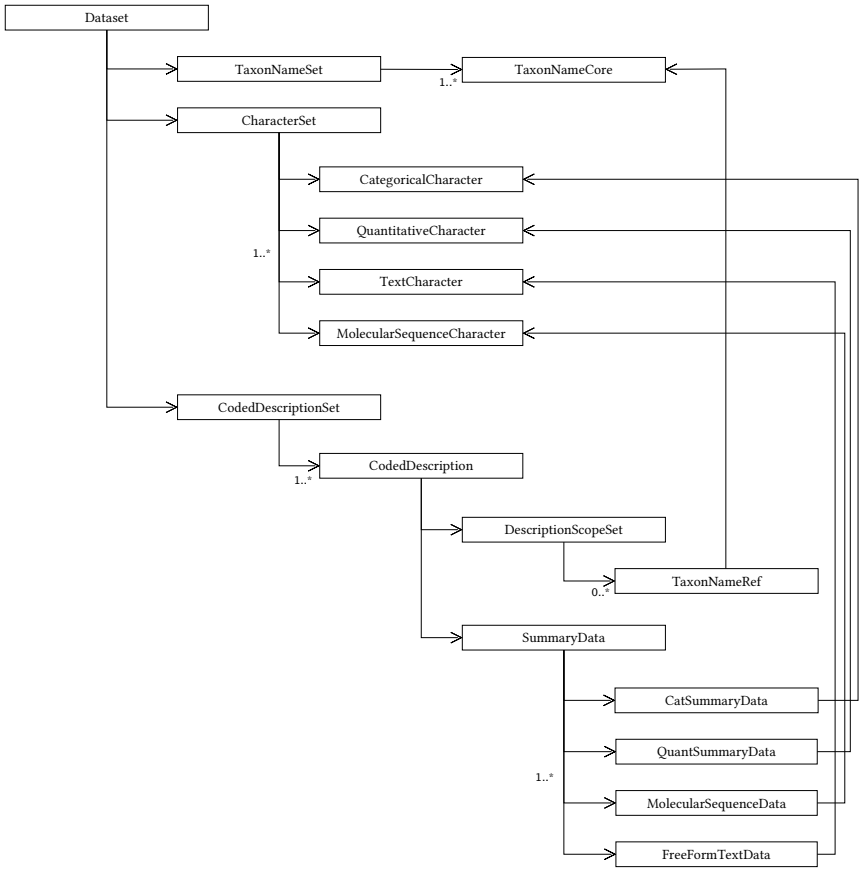

Abbildung 3.5.: Ausschnitt aus dem SDD-Datenmodell Version 1.1 (vereinfacht)

```
<?xml version="1.0" encoding="UTF-8"?>
<Datasets xsi:schemaLocation="http://rs.tdwg.org/UBIF/2006/
    http://rs.tdwg.org/UBIF/2006/Schema/1.1/SDD.xsd"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xmlns="http://rs.tdwg.org/UBIF/2006/">
  ...
  <Dataset>
  ...
  </Dataset>
  ...
</Datasets>
```

Abbildung 3.6.: Wurzelelement eines SDD-Dokumentes

```
<TaxonNames>
  <TaxonName id="tg">
    <Representation>
      <Label>Erythroneura aclys McAtee</Label>
    </Representation>
  </TaxonName>
  <TaxonName id="ri">
    <Representation>
      <Label>Erythroneura acuticephala Robinson</Label>
    </Representation>
  </TaxonName>
  ...
</TaxonNames>
```

Abbildung 3.7.: Liste der Taxa im TaxonNames-Element

MolecularSequenceCharacter und *TextCharacter*. Innerhalb dieser Typen werden Informationen für die Visualisierung gespeichert und der Beschreibungsraum Δ_j wird definiert. Für das System zur Bestimmung von Mikroorganismen im Rahmen des Verfahrens zur Berechnung des Saprobienindex sind nur die Typen CategoricalCharacter und QuantitativeCharacter nutzbar. Auch die Character-Elemente haben, ebenso wie die Elemente des Wertebereichs quantitativer Variablen (States), ein *id*-Attribut, über das sie referenziert werden können.

CategoricalCharacter. CategoricalCharacter beschreiben Merkmale mit einem nominalen oder ordinalen Wertebereich (qualitative Daten). Dazu enthalten sie alle möglichen Merkmalsausprägungen (StateDefinition) in Form einer geordneten Liste (CharacterStateSeq).

QuantitativeCharacter. Quantitative Merkmale werden mithilfe des Datentyps QuantitativeCharacter definiert. Dieser enthält Informationen über die Einheit, Skalen, Wertebereiche usw. Außerdem können Intervalle angegeben werden, um numerische Werte auf Kategorien eines CategoricalCharacter abzubilden.

3.6.3. Beschreibungen

CodedDescriptionSet. Im Datentyp *CodedDescription* werden die Merkmalsausprägungen für Taxa gespeichert. Er enthält eine Liste von Referenzen (Scope), die u.a. auf Taxa verweisen können. Andere mögliche Referenzen auf Proben, geographische Bereiche usw. sind für

```
<Characters>
  <CategoricalCharacter id="y">
    <Representation>
      <Label>Ground color of dorsum</Label>
    </Representation>
    <States>
      <StateDefinition id="oa">
        <Representation>
          <Label>yellow or white</Label>
        </Representation>
      </StateDefinition>
      ...
    </States>
  </CategoricalCharacter>
  ...
  <QuantitativeCharacter id="C">
    <Representation><Label>Body size</Label></Representation>
    <MeasurementUnit><Label>mm</Label></MeasurementUnit>
  </QuantitativeCharacter>
  ...
</Characters>
```

Abbildung 3.8.: Beschreibung der Merkmale im Characters-Element

```
<CodedDescriptions>
  <CodedDescription id="mg">
    <Scope><TaxonName ref="tg"/></Scope>
    <SummaryData>
      <Categorical ref="y"><State ref="oa"/></Categorical>
      ..
      <Quantitative ref="C">
        <Measure type="UMethLower" value="2.7"/>
        <Measure type="UMethUpper" value="3.1"/>
      </Quantitative>
      ...
    </SummaryData>
  </CodedDescription>
  ...
</CodedDescriptions>
```

Abbildung 3.9.: Beschreibung der Merkmalsausprägungen im CodedDescriptions-Element

das System nicht von Bedeutung. Zusätzlich hat der Datentyp eine Liste von Merkmalsausprägungen (SummaryDataSet), die durch die Typen *CatSummaryData*, *QuantSummaryData*, *MolecularSequenceData* und *FreeFormTextData* beschrieben werden. Wiederum sind hier nur die ersten beiden Typen relevant und werden im Weiteren näher erläutert. Alle diese Typen haben eine Referenz auf das zugehörige Merkmal aus dem CharacterSet.

Referenziert ein CodedDescription-Element ein Taxon, stellt es eine symbolische Beschreibung d_C im Sinne der SDA für das Konzept dieses Taxons dar. Einem Taxon können mehrere CodedDescriptions zugeordnet werden. Dies ermöglicht die gesonderte Beschreibung unterschiedlicher Entwicklungsstadien oder Geschlechter. Die Werte für einzelne Merkmalsausprägungen können als Event der SDA angesehen werden. Die Konjunktion aller Events innerhalb einer CodedDescription bildet eine Assertion.

CatSummaryData. Die möglichen Ausprägungen eines qualitativen Merkmals eines Taxons werden mit dem Datentyp CatSummaryData verwaltet. Dieser beinhaltet eine geordnete Liste von Referenzen auf StateDefinition-Elemente.

Zudem hat er ein Attribut *statemodel*, das angibt, wie diese Liste zu interpretieren ist. Mögliche Werte für statemodel sind: *OrSet* und *OrSeq* für geordnete oder ungeordnete Listen von Ausprägungen, die oder-verknüpft werden, *AndSet*, *AndSeq* und *WithSeq* für geordnete oder ungeordnete Listen, die und-verknüpft werden und *Between*, für die Angabe eines Intervalls

Skala	Typ	Merkmal	Merkmalsausprägung
nominal	CategoricalCharacter	Anordnung der Tentakel	{„gebündelt“}
kardinal	QuantitativeCharacter	Anzahl der Makronukleus-Teile	2
kardinal	QuantitativeCharacter	Länge	[10,60]

Tabelle 3.1.: Beispiele für Merkmale und Merkmalsausprägungen

von Ausprägungen, das durch die Angabe der ersten und letzten Ausprägung beschrieben wird. Standardwert für das statemodel-Attribut ist OrSet.

Die Menge von Ausprägungen kann als mehrwertige qualitative symbolische Variable verstanden werden. Außerdem gibt das statemodel-Attribut Hinweise auf die Vergleichsrelation R_j.

QuantSummaryData. Quantitative Merkmalsausprägungen werden mit dem Typ QuantSummaryData angegeben. Im Gegensatz zu der Liste der Ausprägungen im CatSummaryData-Typ, hat dieser Typ Measure- oder PMeasure-Elemente (für parametrisierte Werte) die Informationen über univariate, statistische Messungen enthalten, so z. B. untere und obere Grenzen, Mittelwerte und andere Größen der deskriptiven Statistik sowie Verteilungsparameter. Diese Daten werden durch quantitative symbolische Variablen ausgedrückt.

Modifier und Status. Sowohl der CatSummaryData-Typ als auch der QuantSummaryData-Typ können verschiedene Modifier, z. B. für Häufigkeit oder (Un)Sicherheit, speichern. Im Falle von CatSummaryData handelt es sich dann um eine modale qualitative symbolische Variable. Außerdem können die Datentypen für Merkmalsausprägungen eine Liste von Statusinformationen enthalten, die z. B. angeben, dass ein Merkmal für das referenzierte Taxon nicht anwendbar ist. Diese Statusinformationen können die Werte *ToBeChecked*, *ToBeIgnored*, *NotApplicable*, *DataUnavailable*, *NotInterpretable* oder *DataWithheld* annehmen.

3.6.4. Systeme mit SDD-Support

Verschiedene Softwaresysteme für die Verwaltung taxonomischer Daten bieten Schnittstellen für den SDD-Import bzw. -Export. Dazu gehören z. B. Lucid3[22], Xper2[23] und CDM[24]-basierte Systeme, wie z. B. Taxononic Editor[25].

[22] http://www.lucidcentral.org
[23] http://lis-upmc.snv.jussieu.fr/lis/?q=ressources/logiciels/xper2
[24] http://wp5.e-taxonomy.eu/cdmlib/
[25] http://wp5.e-taxonomy.eu/taxeditor/

4. Anforderungen

Aus der Wissensbasis soll ein interaktiver Bestimmungsschlüssel generiert werden. Die beobachteten Merkmale des zu untersuchenden Mikroorganismuses werden sukzessive vom Benutzer abgefragt. Beginnend mit der Menge aller bekannten Taxa werden schrittweise diejenigen Taxa aus der Ergebnismenge entfernt, deren Beschreibung den beobachteten Merkmalen widerspricht.

4.1. Wissensrepräsentation

Die Zuordnung von Merkmalsausprägungen zu Taxa, wie sie der Fachliteratur zu entnehmen ist, erfolgt einerseits induktiv auf Basis empirischer Untersuchungen. Andererseits begründen bestimmte Merkmale und ihre unterschiedliche Merkmalsausprägungen gerade die Einteilung in die verschiedenen Taxa, so dass diese Zuordnung auch normativen Charakter hat.

4.1.1. Datenmodell

Das Wissen über Eigenschaften von Vertretern bestimmter Taxa wird durch die Konzepte Taxon, Merkmal und Merkmalsausprägung sowie die Beziehungen zwischen diesen Konzepten beschrieben.

A1.1. Der Standard *Structured Descriptive Data (SDD)*[1] legt ein Schema für die Beschreibung dieser Daten fest. Diverse Softwaresysteme für die Verwaltung taxonomischer Daten und die Erzeugung interaktiver Bestimmungsschlüssel verfügen über eine Import- bzw. Export-Schnittstelle für SDD. Daher ist es sinnvoll, das SDD-Datenmodell zu verwenden. Ein Vorteil ist, dass der Wissenserwerb, also die Pflege der Wissensbasis, mit vorhandenen Werkzeugen durchgeführt werden kann. Außerdem wird die Wiederverwendbarkeit der zu erstellenden Komponente dadurch gefördert.

[1] http://www.tdwg.org/standards/116/

4.1.2. Inferenzregeln

Die grundlegende Regel, die sich aus dem Wissen über die Merkmale von Taxa ergibt, lautet *Wenn <Taxon>, dann <Merkmalsausprägungen>* $L \rightarrow M$. Da nicht das gesuchte Taxon, sondern die beobachteten Merkmale vorliegen, lassen sich diese Regeln nicht direkt anwenden. Also ist für alle Taxa eine Verdachtshypothese $(L \rightarrow M) \wedge \neg M \rightarrow \neg L$ (Modus tollens) zu überprüfen.

A1.2. Es sind alle plausiblen Lösungen gesucht. Daher ergibt sich die Lösungsmenge C aus allen Taxa, deren Beschreibung den beobachteten Merkmalen nicht widerspricht. Mehrfachdiagnosen sind dabei ausdrücklich zugelassen, d. h. bei n bekannten Taxa beträgt die Größe des Lösungsraums 2^n.

4.2. Problemlösungskomponente

Im Weiteren werden die, zum Teil aus dem Abschnitt zur symbolischen Datenanalyse bekannten, Bezeichner wie folgt verwendet:

Y_j ist das j-te Merkmal aus den p bekannten Merkmalen $Y_1, \ldots, Y_p$.
SDD: CharacterSet, CategoricalCharacter, QuantitativeCharacter

O_j ist der Wertebereich des j-ten Merkmals.
SDD: CategoricalCharacter, QuantitativeCharacter

d_j bezeichnet die symbolische Beschreibung des j-ten Merkmals aus D_j für ein bestimmtes Taxon.
SDD: CodedDescription

R_j ist die Vergleichsrelation für d_j.

$Y_j(\omega)$ sind die bei einem Mikroorganismus ω beobachtete Ausprägungen des des j-ten Merkmals. Diese werden vom Anwender erfasst.

Das SDD-Datenmodell stellt die zusätzlichen Attribute für Beschreibungen bereit:

$statemodel(d_j)$ bezeichnet die Art der Verknüpfung mehrwertiger quantitativer Variablen mit
$statemodel(d_j) \in \{OrSet, OrSeq, Between, AndSet, AndSeq, WithSeq\}$.
SDD: Categorical/@statemodel (CatSummaryData)

$status(d_j)$ beinhaltet die Statusangaben für die Beschreibung eines bestimmten Merkmals mit
$status(d_j) \subseteq \{ToBeChecked, ToBeIgnored, DataUnavailable, NotInterpretable, DataWithheld, NotApplicable\}$.
SDD: Categorical/Status (CatSummaryData), Quantitative/Status (QuantSummaryData)

$frequency(d_j)$, $certainty(d_j)$ sind Evidenzwerte, die als Tupel $e = (lower, upper)$ für quantitative Variablen bzw. als Menge von Tupeln $\{e_1, \cdots, e_m\}$ für qualitative Variablen mit $m = |d_j|$ angegeben werden.

Zur Laufzeit werden außerdem die Lösungsmenge und weitere Informationen über Merkmale gespeichert:

C ist die Menge der verbleibenden Verdachtshypothesen.

$recorded(Y_j)$ markiert bereits erfasste Merkmale: $recorded : Y \rightarrow \{wahr, falsch\}$.

$observable(Y_j)$ gibt an, ob der Nutzer in der Lage ist, das Merkmal Y_j bei dem betrachteten Mikroorganismus zu untersuchen, oder ob es von der Bewertung ausgenommen werden soll: $observable : Y \rightarrow \{wahr, falsch\}$. Der Standardwert ist $wahr$.

$dp(Y_j, C)$ ist ein Wert für die Trennschärfe[2], der die Fähigkeit des Merkmals Y_j beziffert, die Taxa C anhand ihrer Beschreibungen zu unterscheiden.

A2.1. Um Unsicherheiten bei der Beobachtung abzubilden, kann für jedes Merkmal Y_j eine Menge von Ausprägungen $Y_j(\omega) \in \Delta_j$, $\Delta_j \equiv \mathcal{P}(O_j)$ erfasst werden. Da auch das SDD-Datenmodell mehrwertige Variablen berücksichtigt und auch einwertige Variablen oder Intervalle auf Mengen zurückgeführt werden können, gilt $D_j \equiv \Delta_j$ und $Y_j(\omega), d_j \in \mathcal{P}(O_j)$.

A2.2. Ist ein Anwender in Bezug auf ein bestimmtes Merkmal Y_j völlig unsicher, z. B. wenn zur Beobachtung nötige technische Mittel fehlen, muss die Möglichkeit gegeben sein, dass er dieses Merkmal von der Bewertung ganz ausschließt, so dass dann gilt: $observable(Y_j) = falsch$.

4.2.1. Evidenzen

Das SDD-Datenmodell berücksichtigt Beschreibungen für quantitative und (modale) qualitative Merkmale. Jedem Wert für ein quantitatives Merkmal und jeder möglichen Ausprägung

[2] discriminant power.

eines qualitativen Merkmals können durch Modifier-Elemente Evidenzwerte zugeordnet werden. Wie bereits dargelegt, lassen sich aus diesen Evidenzwerten jedoch keine vergleichbaren Bewertungen der Lösungen ableiten. Somit sind Merkmale mit unsicherer Evidenz entweder als sicher oder als unbekannt zu betrachten. Dafür wird ein weiteres Prädikat $evident(d_j)$

$$evident : D_j \rightarrow \{wahr, falsch\} \tag{4.1}$$

definiert.

A2.3. Für die Bewertung der Evidenz gelten folgende Anforderungen:

1. Einige Merkmale sind bei bestimmten Taxa nicht anwendbar[3]. Dieser Umstand wird durch die Statusangabe $status(d_j) \ni NotApplicable$ bezeichnet. Beschreibungen, die ein bestimmtes Merkmal als nicht anwendbar für ein Taxon kennzeichnen, gelten als evident.

2. Beschreibungen mit bestimmten Statusangaben $status(d_j) \cap \{ToBeChecked, ToBeIgnored, DataUnavailable, NotInterpretable, DataWithheld\} \neq \emptyset$ gelten als nicht evident.

3. Ist ein Merkmal für ein Taxon nicht beschrieben $d_j = \emptyset$, gilt es als nicht evident, es sei denn, es ist als nicht anwendbar gekennzeichnet.

4. Sind keine Evidenzwerte angegeben, gilt das Merkmal als sicher beschrieben.

5. Um unsichere Merkmale mit hoher Evidenz trotzdem für die Bestimmung verwenden zu können, soll über einen Schwellenwert $threshold$ konfiguriert werden, welche Evidenzen als sicher anzusehen sind.

6. Evidenzwerte einer mehrwertigen qualitativen Variable sind dabei zu addieren.

7. Wenn sowohl Werte für $certainty(d_j)$ als auch $frequency(d_j)$ gegeben sind, wird zunächst $certainty(d_j)$ ausgewertet. Diese Auswahl ist willkürlich, genauso wie die Verwendung des oberen Limits.

[3] Z. B. Merkmale von Wimpern bei Taxa, die nicht zu den Wimperntierchen gehören.

$$evident(d_j) := \begin{cases} wahr & \text{, wenn } status(d_j) \ni NotApplicable \\ falsch & \text{, wenn } status(d_j) \neq \emptyset \\ falsch & \text{, wenn } d_j \equiv \emptyset \\ wahr & \text{, wenn } certainty(d_j) = frequency(d_j) = \emptyset \\ wahr & \text{, wenn } \sum_{c \in certainty(d_j)} upper(c) \geq threshold \\ wahr & \text{, wenn } \sum_{f \in frequency(d_j)} upper(f) \geq threshold \\ falsch & \text{, sonst} \end{cases} \tag{4.2}$$

Die Angaben für quantitative Variablen werden dabei als einwertige Mengen $\{(lower, upper)\}$ betrachtet.

4.2.2. Diagnosemodus

Die Bestimmung des Taxons des untersuchten Mikroorganismus geschieht durch die schrittweise Bewertung der hypothetischen Lösungen durch Tests, die die beobachteten Ausprägungen eines Merkmals mit den Beschreibungen der Taxa vergleichen.

Die SDD-Daten beschreiben das Konzept jedes enthaltenen Taxons durch ein symbolisches Object $S_C = (a_C, R_C, d_C)$, das insbesondere eine Assertion der einzelnen Variablen darstellt. Die Relation R_C ist mithin eine Konjunktion der einzelnen Relationen R_j. Durch die Auflösung der Evidenzen mittels Schwellenwert hat die Zuordnungsfunktion die Form $a_C : \Omega \rightarrow \{wahr, falsch\}$.

Die Relationen R_j ergeben sich aus dem Variablentyp, der Evidenz der Beschreibung und dem *statemodel*-Attribut bei qualitativen Variablen. Für den speziellen Anwendungsfall ist es ausreichend, ein- und mehrwertige sowie Intervallvariablen zu berücksichtigen, da symbolische Variablen in Form von Funktionen, Verteilungen oder Histogrammen der zugrundeliegenden Fachliteratur nicht zu entnehmen sind und im Zuge der Auflösung von Evidenzen ohnehin umgewandelt werden müssten.

A2.4. Unter der Berücksichtigung der Regeln für die Evidenz ergeben sich die Relationen R_j für die einzelnen Beschreibungen d_j aus den folgenden Anforderungen:

1. Noch nicht erfasste und nicht evidente Merkmale führen niemals zu einem Widerspruch, so dass gilt: $R_j(Y_j(\omega), d_j) := true$.

2. Bei nicht anwendbaren Merkmalen ist R_j gegeben durch: $Y_j(\omega) \equiv \emptyset$, so dass jede beobachtete Ausprägung einen Widerspruch erzeugt.

3. Für und-verknüpfte mehrwertige qualitative Variablen wird R_j definiert durch: $Y_j(\omega) \supseteq d_j$.

4. Für alle anderen Fälle wird R_j bestimmt durch: $Y_j(\omega) \cap d_j \neq \emptyset$.

$$R_j(Y_j(\omega), d_j) := \begin{cases} wahr & \text{, wenn} \quad \neg recorded(Y_j) \vee \neg evident(d_j) \\ Y_j(\omega) \equiv \emptyset & \text{, wenn} \quad status(d_j) \ni NotApplicable \\ Y_j(\omega) \supseteq d_j & \text{, wenn} \quad statemodel(d_j) \in \{AndSet, AndSeq, WithSeq\} \\ Y_j(\omega) \cap d_j \neq \emptyset & \text{, sonst} \end{cases} \tag{4.3}$$

A2.5. Die Hypothese, dass der beobachtete Mikroorganismus zu dem durch das betrachtete symbolische Objekt $S_C = (a_C, R_C, d_C)$ beschriebene Taxon gehört, wird dann insgesamt durch

$$a_c(\omega) = \bigwedge_{j=1}^{p} R_j(Y_j(\omega), d_j) \tag{4.4}$$

widerlegt oder bestätigt.

A2.6. Das System soll eine möglichst sichere und effiziente Bestimmung ermöglichen. Die Reihenfolge der abgefragten Merkmale soll so gewählt werden, dass die Anzahl der zur Bestimmung benötigten Merkmale, also die Tiefe des Entscheidungsbaumes, minimiert wird und damit auch die Fehlerwahrscheinlichkeit sinkt. Dazu bewertet das System die zur Verfügung stehenden Merkmale mit $observable(Y_j) \wedge \neg recorded(Y_j)$ anhand ihrer Fähigkeit, die Taxa voneinander zu unterscheiden.

Sei $C = \{C_1, \cdots C_t\}$ die Menge der t Konzepte (Taxa), die durch die symbolischen Objekte $S_{C_1} = (a_{C_i}, R_{C_i}, d_{C_i}), 1 \leq i \leq t$ werden. Dann kann eine Funktion $R'_j : D_j \rightarrow \mathcal{P}(C)$ definiert werden, die allen möglichen Beschreibungen $d'_j \in D_j$ eines Merkmals Y_j die Menge der Konzepte zuordnet, die der Beschreibung entsprechen:

$$R'_j(d'_j) := \{C_i \mid R_{ji}(d'_j, d_{ji}) = wahr\} \tag{4.5}$$

Die Menge C'_j der so erzeugten Mengen

$$C'_j := \{R'_j(d'_j) \mid d'_j \in D_j\} \tag{4.6}$$

bildet eine Überdeckung von C, da für jedes Paar R_{ji}, d_{ji} ein d'_j so gewählt werden kann, dass $R_{ji}(d'_j, d_{ji})$ wahr wird (siehe 4.3).

Um die Trennschärfe des Merkmals Y_j zu bewerten, ist C'_j zu untersuchen. C'_j hat dabei folgende wünschenswerte Eigenschaften:

1. Die Elemente von C'_j sind gleich mächtig.
   In dem Sonderfall, dass einige Elemente von C'_j der gesamten Lösungsmenge C entsprechen, während alle anderen leer sind, sind die Konzepte anhand dieses Merkmals nicht zu unterscheiden. Enthalten alle Elemente von C'_j höchstens eine Lösung, stellt Y_j eine eindeutige Abbildung dar, so dass anhand einer Ausprägung direkt auf eine mögliche Lösung geschlossen werden kann. Die Trennschärfe ist in dem Fall maximal.

2. Die Elemente von C'_j unterscheiden sich voneinander.
   D. h. die erwartete Schnittmenge $A \cap B$ mit $A, B \in C'_j, A \neq B$ zweier zufällig gewählter Elemente ist möglichst klein.

Die Funktion $dp : (Y_j, C) \to \mathbb{R}_{\geq 0}$ ist also so zu wählen, das ihre Funktionswerte umso größer sind, je mehr C'_j diesen Anforderungen entspricht. Insbesondere soll $dp(Y_j, C) = 0$ gelten, wenn die Taxa durch dieses Merkmal nicht unterscheidbar sind.

A2.7. Verbleibende Hypothesen werden solange anhand neuer Merkmale getestet, bis keine weiteren Merkmale zur Unterscheidung zur Verfügung stehen oder die Menge der möglichen Lösungen weniger als zwei Elemente enthält.
Nachbedingungen:

- $|\{Y_j \mid observable(Y_j) \wedge \neg recorded(Y_j) \wedge dp(Y_j) > 0\}| = 0 \vee |\{S_C \mid a_c(\omega)\}| < 2$

4.2.3. Konsistenzmodus

A2.8. Wenn eine Menge von Merkmalen erfasst wurde, die zu allen Beschreibungen im Widerspruch steht, die Lösungsmenge also leer ist, soll das System auf Anforderung die am we-

nigsten widersprüchlichen Lösungen ausgeben. Der dabei zu minimierende Wert ist die Anzahl der nicht übereinstimmenden Merkmale. Also wird $difference : \Delta \times D \to \mathbb{N}$ definiert als

$$difference(Y(\omega), d_C) = |\{Y_j \mid 1 \leq j \leq p \wedge \neg R_j(Y_j(\omega), d_j)\}| \,. \tag{4.7}$$

4.3. Nutzerschnittstelle

4.3.1. Auswahl der beobachteten Merkmale

A3.1. Bei der Diagnose wird das für die Unterscheidung der Taxa nach A2.6 bestbewertete Merkmal bei jedem Schritt vorausgewählt.

A3.2. Zudem soll der Nutzer die Möglichkeit haben, ein anderes Merkmal auszuwählen. Dazu ist eine Auswahlliste anzuzeigen, die die Merkmale mit $observable(Y_j) \wedge \neg recorded(Y_j) \wedge dp(Y_j) > 0$ enthält.

A3.3. Die Nutzerschnittstelle muss laut A2.2 die Möglichkeit bieten, ein ausgewähltes, noch nicht erfasstes Merkmal als „nicht erkennbar“ zu kennzeichnen. Bei einer leeren Auswahl oder der Auswahl aller Möglichkeiten würden ansonsten möglicherweise durch 4.3 Lösungen fälschlich ausgeschlossen.

4.3.2. Eingabe eines beobachteten Merkmals

A3.4. Nachdem ein Merkmal Y_j durch den Nutzer oder automatisch ausgewählt wurde, wird abhängig vom Variablentyp des Merkmals ein Formular für die beobachteten Ausprägungen dargestellt.

Bei qualitativen Variablen besteht dieses aus einer Auswahlliste der möglichen Merkmalsausprägungen O_j mit der Möglichkeit zur Mehrfachauswahl, um A2.1 zu gewährleisten.

Für quantitative Variablen ist ein Eingabefeld für einen numerischen Wert oder ein Intervall anzuzeigen. Alternativ kann das Intervall zwischen dem kleinsten und größten Wert des Merkmals aller Beschreibungen nutzerfreundlich partitioniert werden, z. B. durch Intervalle mit Grenzen in Form von Vielfachen von 10er-Potenzen bzw. $\frac{1}{2}$-, $\frac{1}{4}$- oder $\frac{1}{5}$-Bruchteilen davon, und als Auswahlliste dargestellt werden.

Der Nutzer trifft dann eine Auswahl und übernimmt diese für die weitere Bewertung der möglichen Lösungen.

4.3.3. Einbinden von Medien

A3.5. Um den Nutzer bei der Eingabe der beobachteten Merkmale und der Verifizierung der Lösungsmenge zu unterstützen, sollen die mit Taxa und Merkmalsausprägungen verknüpften Bild- und Videodaten angezeigt werden.

4.4. Pflege der Wissensbasis

Im Rahmen des Knowledge Engineering sind die Beschreibungen der Taxa in den von der DIN-Norm referenzierten Fachtexten in abstrakte Merkmale und deren mögliche Ausprägungen zu überführen. Dies kann mithilfe von Bestimmungssoftware mit einer SDD-Exportschnittstelle, wie z. B. *Xper2*[4] erfolgen. Die so erzeugte Wissensbasis wird dem System als XML-Dokument nach dem SDD-Schema zur Verfügung gestellt.

[4] http://lis-upmc.snv.jussieu.fr/lis/?q=en/resources/software/xper2

5. Datenmodelle und Algorithmen

5.1. Wissensbasis

5.1.1. Datenmodell

Für die Abbildung der Wissensbasis werden folgende Klassen und Schnittstellen definiert (siehe Abb. 5.1):

Concept: Die bekannten Taxa bzw. Stadien derselben werden durch die Klasse `Concept` modelliert. Objekte dieser Klasse bilden die möglichen Lösungen der Klassifikation (SDA: S_C, SDD: TaxonName).

***Character*:** Mit der Schnittstelle `Character` und ihren Implementierungen `QuantitativeCharacter` und `CategoricalCharacter` werden die Merkmale abgebildet (SDA: Y_j, SDD: QuantitativeCharacter, CategoricalCharacter).

StateDefinition: Die Klasse `StateDefinition` verwaltet Elemente des Wertebereichs eines qualitativen (kategoriellen) Merkmals (SDA: Elemente von O_j, SDD: StateDefinition).

Description: Diese Klasse erfasst die Beschreibungen der Konzepte (SDA: d_C, SDD: CodedDescription). Die Methode `match` bildet die Zugehörigkeitsfunktion a_C.

***CharacterDescription*:** Beschreibungen einzelner Merkmale werden durch die Klassen `Categorical` und `Quantitative` erfasst, die diese Schnittstelle implementieren (SDD: Quantitative, Categorical). Die Relation R_j wird durch die Methode `match` implementiert.

***SymbolicSet*:** Wie in Anforderung A2.1 dargelegt, liegen die Beschreibungen eines Merkmals und dessen beobachtete Werte im gleichen Raum $\mathcal{P}(O_j)$. `SymbolicSet` ist eine abstrakte Repräsentation dieses Raumes, die die in 4.3 definierten Vergleichsoperationen bereitstellt. `NumericSymbolicSet` und `CategoricalSymbolicSet` implementieren diese Schnittstelle (SDA: d_j und $Y_j(\omega)$).

Die Objekte dieser Klassen sind durch Parsen der SDD-Daten im XML-Format zu instanziiern. Nach der Instanziierung sind diese Objekte unveränderlich.

5.1.2. Bewertung der Evidenz

Die Bewertung der Evidenz wird ausschließlich für die `match`-Methode einer Merkmalsbeschreibung in der Schnittstelle `CharacterDescription` verwendet. Die entsprechenden Methoden können daher nicht öffentlich implementiert werden. Listing 5.1 zeigt beispielhaft einen Auszug aus einer Java-Implementierung der `Quantitative`-Klasse und Listing 5.2 ein Beispiel für die `Categorical`-Klasse.

```
public class Quantitative implements CharacterDescription {
  private Character character;
  private Set<Status> status;
  private NumericSymbolicSet range;
  private Evidence certainty;
  private Evidence frequency;
  ...
  private boolean evident(float threshold) {
    return (status.contains(Status.NotApplicable)) ?
        true
        : (!status.isEmpty()) ?
            false :
            : (range==null) ?
                false
                : (certainty==null) ?
                    (frequency==null) ?
                        true : frequency.upper >= threshold
                    : certainty.upper >= threshold;
  }

  public boolean match(SymbolicSet other, float threshold) {
    assert(other instanceof NumericSymbolicSet);
    return evident(threshold) ?
        !range.intersection(other).empty() : true;
  }
}
```

Listing 5.1: Beispiel für eine Java-Implementierung der Quantitative-Klasse

```
public class Categorical implements CharacterDescription {
```

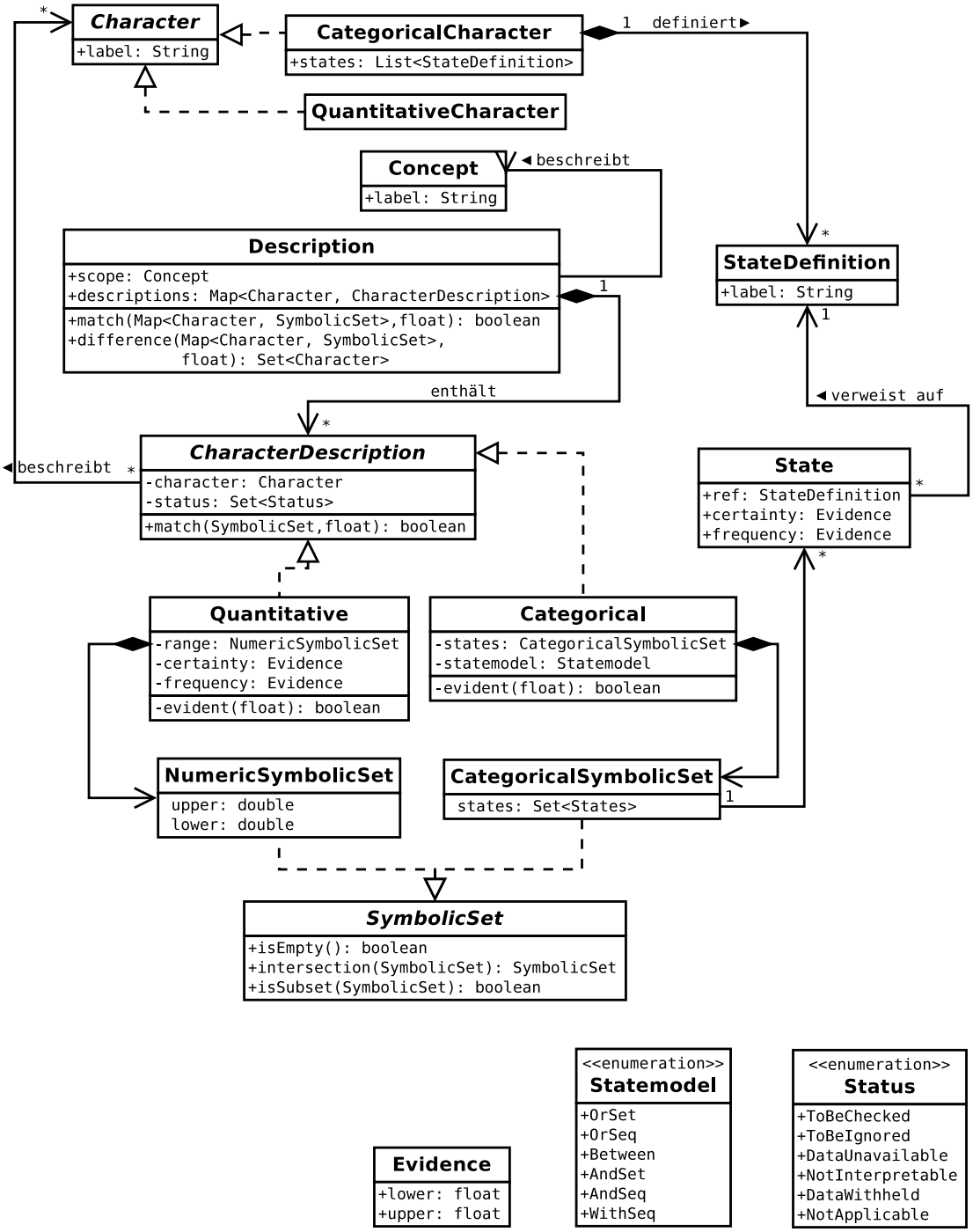

Abbildung 5.1.: UML-Datenmodell der Wissensbasis

```
  private Character character;
  private Set<Status> status;
  private CategoricalSymbolicSet states;
  private Statemodel statemodel;

  private float sumCertaintyUpper;  // diese Variablen könnten im
  private float sumFrequencyUpper;  // Konstruktor berechnet werden
  ...
  private boolean evident(float threshold) {
    return (status.contains(Status.NotApplicable)) ?
        true
        : (!status.isEmpty()) ?
            false
            : (states==null) ?
                false
                : (sumCertaintyUpper == 0) ?
                    (sumFrequencyUpper == 0) ?
                        true : sumFrequencyUpper >= threshold
                    : sumCertaintyUpper >= threshold;
  }

  public boolean match(SymbolicSet other, float threshold) {
    assert(other instanceof CategoricalSymbolicSet);
    return !evident(threshold) ?
        true
        : status.contains(Status.NotApplicable) ?
            other.isEmpty()
            : (statemodel==Statemodel.AndSet
                    || statemodel==Statemodel.AndSeq
                    || statemodel==Statemodel.WithSeq) ?
                states.isSubset(other)
                : !states.intersection(other).empty();
  }
}
```

Listing 5.2: Beispiel für eine Java-Implementierung der Categorical-Klasse

5.1.3. Vergleichsfunktionen

Die `match`-Methoden der `Quantitative`- und `Categorical`-Klasse sind die Umsetzung der Konditionen in der R_j-Relation 4.3. Die weitere Vergleichslogik ist in den Methoden der Implementierungen von `SymbolicSet` enthalten.

```
public class NumericSymbolicSet implements SymbolicSet {
  double lower;
  double upper;

  public NumericSymbolicSet(double lower, double upper) {
    this.lower = lower;
    this.upper = upper;
  }

  public boolean isEmpty() {
    return lower > upper;
  }

  public SymbolicSet intersection(SymbolicSet other) {
    assert(other instanceof NumericSymbolicSet);
    return new NumericSymbolicSet(Math.max(this.lower, other.lower,
        Math.min(this.upper, other.upper)));
  }

  public boolean isSubset(SymbolicSet other) {
    assert(other instanceof NumericSymbolicSet);
    return other.lower<=this.lower && other.upper>=this.upper;
  }
}
```

Listing 5.3: Java-Implementierung der NumericSymbolicSet-Klasse

```
public class CategoricalSymbolicSet implements SymbolicSet {
  Set<States> states;

  public CategoricalSymbolicSet(Set<States> states) {
    this.states = states;
  }

  public boolean isEmpty() {
```

```
9       return states.isEmpty();
10    }
11
12    public SymbolicSet intersection(SymbolicSet other) {
13      assert(other instanceof CategoricalSymbolicSet);
14      return new CategoricalSymbolicSetstates.retainAll(other.states));
15    }
16
17    public boolean isSubset(SymbolicSet other) {
18      assert(other instanceof CategoricalSymbolicSet);
19      return other.containsAll(this.states);
20    }
21  }
```

Listing 5.4: Java-Implementierung der CategoricalSymbolicSet-Klasse

5.2. Diagnosemodus

Hauptaufgabe der Komponente ist die Suche möglicher Lösungen im Diagnosemodus. Dazu wird eine Methode zur sicheren Klassifikation verwendet, die ausgehend von der Menge aller möglichen Lösungen, also Taxa, die durch das jeweils ausgewählte Merkmal und dessen beobachtete Ausprägungen vorgegebene Regel auswertet, und die Lösungsmenge so durch Rückwärtsverkettung reduziert. Diese schrittweise Reduktion mit beliebiger Reihenfolge der Merkmale ist durch die Assoziativität und Kommutativität der in der a_C-Relation 4.4 verwendeten Konjunktion möglich.

5.2.1. Datenmodell

Für Arbeitsspeicher des Diagnosemodus werden die folgenden Klassen definiert (siehe Abb. 5.2):

Observation: Objekte dieser Klasse verwalten die erfassten Merkmale und deren beobachtete Werte (SDA: $Y_j(\omega)$).

Identification: Die Klasse Identification bildet den Arbeitsspeicher. Sie enthält die Wissensbasis in Form der Beschreibungen und verfügbaren Merkmale, sowie die erfassten Merkmale und Zwischenergebnisse. Die Klasse hat die Eigenschaften:

- `descriptions` ist die unveränderliche Menge der Beschreibungen aus der Wissensbasis.

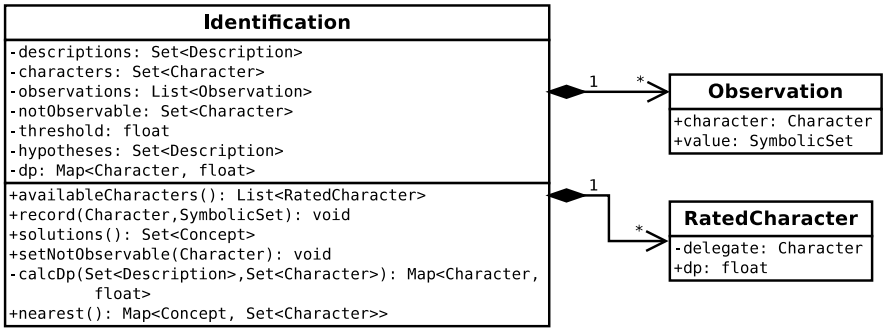

Abbildung 5.2.: UML-Datenmodell des Arbeitsspeichers bei der Diagnose

- `characters` enthält die unveränderliche Menge der Merkmale aus der Wissensbasis.
- `observations` speichert die geordnete Liste der Beobachtungen.
- `notObservable` ist die Menge der nicht erkennbaren Merkmale.
- `hypotheses` speichert die Menge der infrage kommenden Beschreibungen und enthält zunächst alle Elemente aus `descriptions`.
- `threshold` ist der Schwellenwert für Evidenzwerte.

RatedCharacter: Eine Wrapperklasse für Character, die den Wert für die Trennschärfe speichert und alle anderen Zugriffe an das Basis-Character-Objekt weiterleitet. Anhand der Trennschärfe ist die `availableCharacters`-Liste zu sortieren.

5.2.2. Ablauf

Die Diagnose läuft vereinfacht wie folgt ab:

1. Initialisierung des Identification-Objektes
2. Berechnung der Trennschärfe der verfügbaren Merkmale
3. Hole die verfügbaren Merkmale
4. Wenn keine Merkmale verfügbar, weiter mit 13.
5. Wähle das bestbewertete Merkmal
6. Anzeige des Formulars für die Eingabe der Beobachtung

7. Wenn das Merkmal nicht erkennbar ist, Merkmal zu `notObservable` hinzufügen, weiter mit 3.
8. Wenn der Nutzer ein anderes Merkmal auswählt, weiter mit 6.
9. Erzeuge ein SymbolicSet aus der Eingabe
10. Filtere `hypotheses` anhand der `match`-Methoden
11. Wenn Anzahl `hypotheses` kleiner 2, weiter mit 13.
12. Weiter mit 2.
13. Anzeige der Elemente in `hypotheses`

Abb. 5.3 zeigt ein detaillierteres Aktivitätsdiagramm der Diagnose.

Dieses Verfahren entspricht der dynamischen Erzeugung eines (Teil-)Entscheidungsbaumes entlang eines, durch die gewählten Merkmale und Ausprägungen bestimmten, Pfades. Die Knoten des Baumes bestehen dabei aus den Merkmalen, die Kanten aus den möglichen Merkmalsausprägungen. Kanten nicht gewählter Merkmalsausprägungen werden dabei nicht weiter beachtet.

Würde dieser Baum vorberechnet werden, entfiele die komplexe Berechnung der Trennschärfe bei jedem Reduktionsschritt. Die Möglichkeit für den Anwender, die Reihenfolge durch Auswahl eines anderen Merkmals zu ändern, erfordert dann aber die Berechnung von bis zu $p!$ Bäumen (p: Anzahl der Merkmale).

Berücksichtigt man außerdem die Möglichkeit zur Verknüpfung, insbesondere zur Konjunktion, von einzelnen Merkmalsausprägungen sowohl in den Beschreibungen als auch bei der Erfassung, wird klar, dass die Bäume auch sehr breit werden können. So hat ein Knoten für das Merkmal Y_j bis zu $|\mathcal{P}(O_j)| = 2^{|O_j|}$ ausgehende Kanten.

Der Aufwand für Berechnung und Speicherung der Bäume erscheint daher übertrieben. Zudem würden diese Berechnungen mit einer Änderung der Wissensbasis hinfällig. Sinnvoll kann jedoch ein Caching der berechneten Trennschärfewerte in der Form $C \rightarrow dp : Y \rightarrow [0, 1]$ (C: Menge der Verdachtshypothesen) sein.

5.2.3. Trennschärfe

Der Testauswahl kommt als Teilaufgabe der Diagnose eine besondere Bedeutung zu. Eine geeignete Auswahl der Tests kann die Anzahl der nachgefragten Merkmale reduzieren und so zur Nutzerfreundlichkeit als auch zur Fehlervermeidung beitragen.

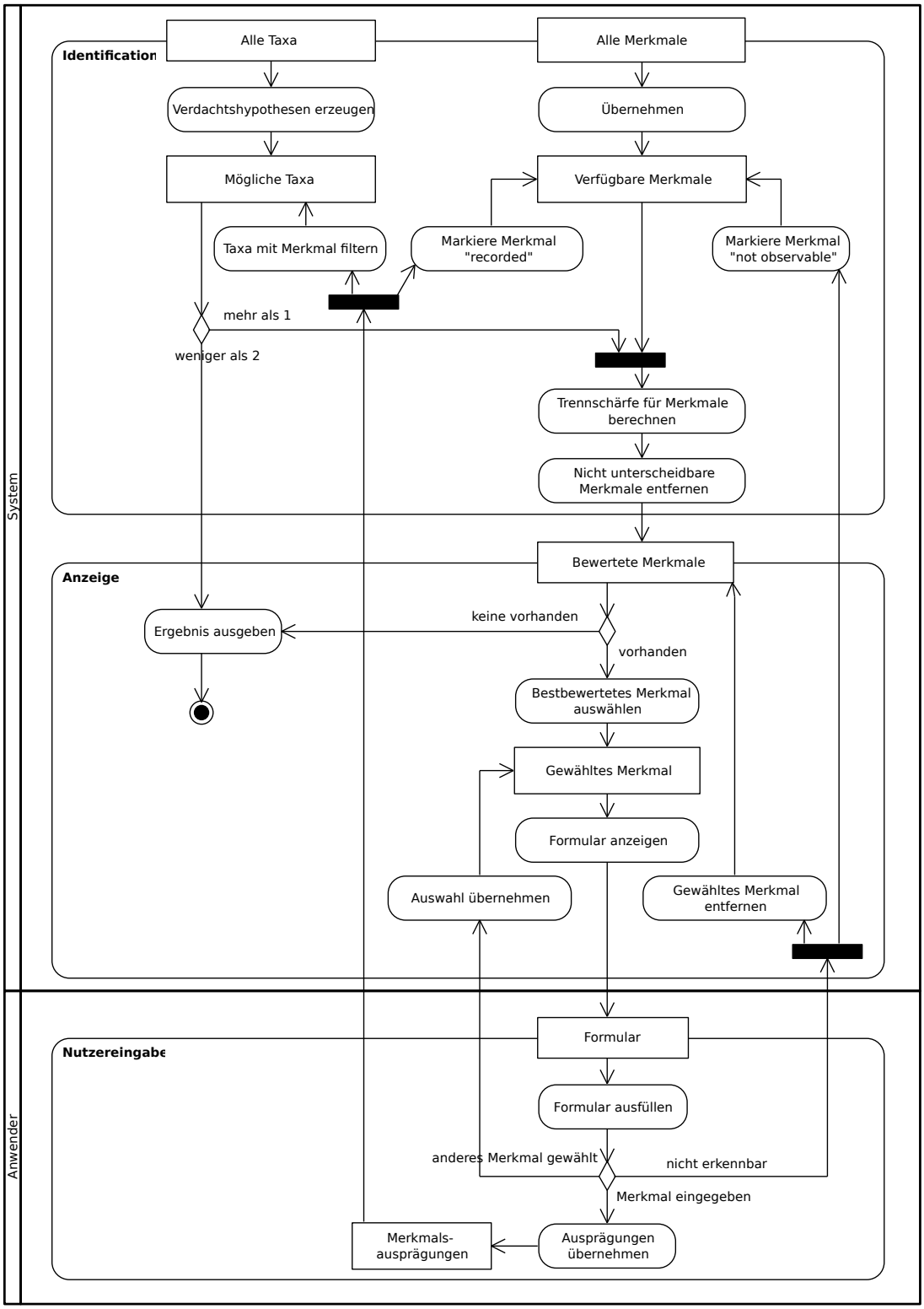

Abbildung 5.3.: Aktivitätsdiagramm der Diagnose

Die Trennschärfe soll dabei ein Maß zur Bewertung eines Merkmals sein, das angibt, inwieweit die Kenntnis des Merkmals zur Verringerung der Unsicherheit in Bezug auf die Lösung beiträgt. Für die Berechnung eines derartigen Maßes existieren verschiedene Ansätze.

Das von Quinlan (1986) entwickelte Verfahren ID3 zur rekursiven Induktion von Entscheidungsbäumen und dessen Nachfolger C4.5 und C5.0 verwenden die *Shannon-Entropie*[1] und den darauf basierenden *Information Gain* für die Bewertung und Auswahl der Attribute. Eine andere Familie von Verfahren basiert auf Unabhängigkeitstests der Variablen Attribut und Lösung (z. B. CHAID[2]).

Diese Verfahren gehen von einer Partitionierung der Lösungsmenge durch die Attribute aus. Da bei dem Diagnoseverfahren jedoch Überdeckungen der Lösungsmenge erzeugt werden, wie in Anforderung A2.6 gezeigt wird, sind die Elemente dieser Überdeckung zu untersuchen.

Unähnlichkeit der möglichen Lösungsmengen. Für die Bewertung der Un-/Ähnlichkeit zweier Mengen A und B sind verschiedene Maße gebräuchlich. Eine Variante ist der *Jaccard-Koeffizient*:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \text{ Sonderfall: } J(\emptyset,\emptyset) = 1 \tag{5.1}$$

für die Ähnlichkeit von Mengen und $1 - J(A,B)$ für die Unähnlichkeit. Xper² bietet eine weitere Funktion zur Bewertung der Unähnlichkeit[3]:

$$dXper(A,B) = \begin{cases} 1 & \text{, wenn } A \cap B = \emptyset \\ 0 & \text{, sonst.} \end{cases} \tag{5.2}$$

Sei also $diff$ eine Funktion $diff : \mathcal{P}(C) \times \mathcal{P}(C) \rightarrow \mathbb{R}_{\geq 0}$ für die Unähnlichkeit von zwei Lösungsmengen, wobei ein Wert von 0 anzeigt, dass die Mengen äquivalent sind. C'_j beschreibt wieder die Menge der vom Merkmal Y_j erzeugten Lösungsmengen. Dann kann der Erwartungswert $diff_E(C'_j)$ der Unähnlichkeit zweier verschiedener Elemente aus C'_j entsprechend der Anforderung A2.6 durch

$$diff_E(C'_j) = \frac{2}{(m-1)m} \sum_{a=1}^{m} \sum_{b>a}^{m} diff(C'_{ja}, C'_{jb}) \text{ mit } m = |C'_j| \tag{5.3}$$

[1] Shannon (2001).
[2] Kass (1980).
[3] http://www.infosyslab.fr/lis/?q=en/resources/software/cai/xper2/documentation_en/FAQ_use#identification.power

berechnet werden. Sind alle Lösungsmengen in C'_j äquivalent, die Taxa sind anhand des betrachteten Merkmals also nicht unterscheidbar, hat $diff_E(C'_j)$ ebenfalls den Wert 0.

Gleiche Mächtigkeit der Lösungsmengen. $diff_E$ wird auch dann maximal, wenn C'_j genau zwei Elemente hat, wovon eines eine leere Menge ist und das andere der gesamten Lösungsmenge entspricht. In diesem Fall sind die Taxa durch das Merkmal jedoch nicht unterscheidbar. Es wird also eine weitere Funktion g gesucht, die für diesen Fall den Wert 0 annimmt, so dass dp durch $diff_E(C'_j) \cdot g(C'_j)$ definiert werden kann. Dabei kann es sich um eine sehr einfache konditionale Funktion handeln, die diesen Sonderfall speziell prüft, und ansonsten 1 zurückgibt.

Alternativ können die Lösungsmengen in C'_j auch als Werte einer Zufallsvariablen X über dem Alphabet Z angesehen werden, wobei Z die Elemente aus C'_j ohne Duplikate enthält. Mit den Häufigkeiten h der Elemente aus Z in C'_j kann dann die Entropie

$$\mathcal{H}(X) = -\sum_{i=1}^{|Z|} \frac{h_i}{|C'_j|} log_{|Z|} \frac{h_i}{|C'_j|} \rightarrow [0,1] \tag{5.4}$$

berechnet werden. Die Entropie hat auch die Eigenschaft, den Wert 0 anzunehmen, falls eine Häufigkeit h_i den Wert $|C'_j|$ hat und alle anderen somit 0 sind. Damit erfüllt sie die oben gestellte Bedingung und stellt außerdem ein Maß für die Vorhersagbarkeit in Bezug auf X dar, das allerdings die Überlappungen von Lösungsmengen nicht berücksichtigt.

Unähnlichkeit der Beschreibungen. Da jedem Konzept, also jeder Lösung, genau eine Beschreibung zugeordnet ist, ist die Unähnlichkeit der Beschreibungen proportional zur Unähnlichkeit der Lösungsmengen. Unter dieser Voraussetzung kann die Trennschärfe eines Merkmals ohne die Berechnung der Lösungsmengen durch Vergleich der Beschreibungen bewertet werden. Dabei ist aber zu beachten, dass, im Gegensatz zu den Lösungsmengen in C'_j, auch unendliche Mengen im Falle quantitativer Variablen berücksichtigt werden müssen. Eine mögliche Funktion für Ähnlichkeit zweier durch Intervalle beschriebener quantitativer Variablen $a = [lower_a, upper_b]$ und $b = [lower_b, upper_b]$ ist:

$$diff(a,b) = \begin{cases} 1, \text{ wenn } min\{upper_a, upper_b\} \leq max\{lower_a, lower_b\} \\ 1 - \frac{min\{upper_a, upper_b\} - max\{lower_a, lower_b\}}{max\{upper_a, upper_b\} - min\{lower_a, lower_b\}}, \text{ sonst.} \end{cases} \tag{5.5}$$

Qualitative Beschreibungen lassen sich mit den o.g. Funktionen vergleichen. Der Erwartungswert der Unähnlichkeit der Beschreibungen für Y_j berechnet sich dann durch:

$$E(diff) = \frac{2}{(m-1)m} \sum_{a=1}^{m} \sum_{b>a}^{m} diff(d_{ja}, d_{jb}) \text{ mit } m = |C|. \tag{5.6}$$

Auch bei dieser Variante ist zusätzlich die Mächtigkeit der Mengen zu berücksichtigen.

5.3. Konsistenzmodus

Die Methode `nearest` der Identification-Klasse berechnet die am wenigsten widersprüchlichen Lösungen, wie in Anforderung A2.8 gefordert. Dazu wird für jedes Description-Objekt die `difference`-Methode verwendet, die die Menge der nicht übereinstimmende Merkmale zurückgibt.

Bei diesem Verfahren handelt es sich um eine einfache Variante der überdeckenden Klassifikation.

6. Zusammenfassung

6.1. Schlussbemerkung

Im Rahmen dieser Arbeit wurden Methoden zur wissensbasierten Bestimmung von Mikroorganismen evaluiert und Möglichkeiten zur Modellierung der aus organismischen Beschreibungsdaten bestehenden Wissensbasis untersucht. Auf dieser Basis wurden die Anforderungen für ein entscheidungsunterstützendes System entwickelt und ein Datenmodell, sowie die wesentlichen Algorithmen entworfen.

6.2. Fazit

Die Bestimmung von Mikroorganismen ist ein klassisches Klassifikationsproblem, für das bereits vielfältige Lösungsmethoden beschrieben worden sind. Bei der Untersuchung der Unsicherheiten in den Beschreibungen von Mikroorganismen wurde klar, dass die Wissensbasis enge Grenzen für die Methoden zum Schließen unter Unsicherheit setzt. Im Ergebnis kann das System nur plausible Lösungen ausgeben, aber nicht zwischen wahrscheinlichen und unwahrscheinlichen Lösungen unterscheiden. Die endgültige Entscheidung muss in jedem Fall dem Nutzer überlassen bleiben.

Mit dem „Structured Descriptive Data“-Format steht ein umfassendes Schema für die Beschreibungsdaten zur Verfügung. Die Verwendung dieses Formates ermöglicht die Wiederverwendung des konzipierten Systems. Zudem hat sich gezeigt, dass die symbolische Datenanalyse die theoretischen Grundlagen für die Struktur der SDD liefert.

Wesentliche Teile des Entwurfs der Bestimmungskomponente liessen sich auf dieser Basis in Form mathematischer Funktionen beschreiben. Die Implementierung in einer funktionalen Programmierungsprache erscheint sinnvoll, insbesondere weil für die Bewertung der Trennschärfe von Merkmalen bei der Diagnose eine große Anzahl von teilweise komplexen Berechnungen erforderlich ist.

A. Saprobienindex nach DIN 38410-1:2004-10

Die DIN 38410-1:2004-10[1] legt ein Verfahren zur Bestimmung des Saprobienindex in Fließgewässern fest. Mit diesem Verfahren wird anhand von Saprobien in einer Probe, die als Bioindikatoren dienen, einen Index zur Bewertung der saprobiellen Belastung des Gewässers ermittelt. Dieser Index ermöglicht eine Einordnung in Gewässergüteklassen (Tabelle A.1).

Saprobien und saprobielle Valenz. Die DIN-Norm definiert Saprobien[2] als Taxa, die aufgrund ihrer saprobiellen Valenz geeignet sind, bestimmte Saprobiebereiche aufzuzeigen. Die saprobielle Valenz bezeichnet dabei die Saprobiebereiche, in denen ein spezielles Taxon existieren kann. Saprobien sind mithin Gruppen von Indikatororganismen für Intensität des oxidativen biologischen Abbaus organischer Substanz. Die saprobiellen Valenzen wurden im Rahmen empirischer Untersuchungen als Histogramm des Auftretens eines bestimmten Taxons in den einzelnen Saprobiebereichen ermittelt.

Saprobiewert und Indikationsgewicht. Aus den saprobiellen Valenzen wird für den jeweiligen Saprobier ein Saprobiewert als dimensionslose Zahl zwischen 1,0 und 4,0 ermittelt. Zusätzlich wird ein Indikationsgewicht als dimensionslose Zahl 1, 2, 4, 8 oder 16 angegeben, welches sich aus der Breite der saprobiellen Valenz ergibt und somit ein Maß für die Indikatorqualität des Saprobiers darstellt. Dabei gilt: Je größer das Indikationsgewicht ist, desto schmaler der Saprobiebereich, in dem der Indikatororganismus existieren kann.

Abundanzziffer. Die Abundanzziffer ist ein Maß für Individuendichte eines Taxons und wird als ganze Zahl zwischen 1 (Einzelfund) und 7 (Massenvorkommen) angegeben.

[1] DIN384101 (2004).
[2] Einzahl: Saprobier

Güte-klasse	Grad der organischen Belastung	Saprobität (Saprobiestufe)	Saprobien-index
I	unbelastet bis sehr gering belastet	Oligosaprobie	1,0 - < 1,5
I - II	gering belastet	Oligosaprobie mit betamesosaprobem Einschlag	1,5 - < 1,8
II	mäßig belastet ausgeglichene	Betamesosaprobie	1,8 - < 2,3
II - III	kritisch belastet	Alpha-beta-mesosaprobie Grenzzone	2,3 - < 2,7
III	stark verschmutzt ausgeprägte	Alphamesosaprobie	2,7 - < 3,2
III - IV	sehr stark verschmutzt	Polysaprobie mit alpha-mesosaprobem Einschlag	3,2 - < 3,5
IV	übermäßig verschmutzt	Polysaprobie	3,5 - < 4,0

Tabelle A.1.: Güteklassen nach Gewässergütekarte der Bundesrepublik Deutschland

Saprobienindex. Der Saprobienindex einer Gewässerprobe wird mit der Formel (A.1)

$$S = \frac{\sum_{i=1}^{n} s_i \cdot A_i \cdot G_i}{\sum_{i=1}^{n} A_i \cdot G_i} \tag{A.1}$$

berechnet, dabei ist

- S der Saprobienindex
- i die laufende Nummer des Taxons
- s_i der Saprobiewert des i-ten Taxons
- A_i die Abundanzziffer des i-ten Taxons
- G_i das Indikationsgewicht des i-ten Taxons
- n Anzahl der Taxa

Literaturverzeichnis

[DIN384101 2004] *DIN 38410-1:2004-10 Deutsche Einheitsverfahren zur Wasser-, Abwasser- und Schlammuntersuchung - Biologisch-ökologische Gewässeruntersuchung (Gruppe M) - Teil 1: Bestimmung des Saprobienindex in Fließgewässern (M 1).* Deutsches Institut für Normung e.V., 2004

[Atkinson und Gammerman 1987] Atkinson, W. D. ; Gammerman, A.: An Application of Expert Systems Technology to Biological Identification. In: *Taxon* 36 (1987), Nr. 4, S. 705–714

[Beierle und Kern-Isberner 2003] Beierle, C. ; Kern-Isberner, G.: *Methoden wissensbasierter Systeme : Grundlagen, Algorithmen, Anwendungen.* 2., überarb. u. erw. Aufl. Braunschweig : Vieweg, 2003

[Borgelt u. a. 2003] Borgelt, C. ; Timm, H. ; Kruse, R.: *Handbuch der Künstlichen Intelligenz.* Kap. 9. Unsicheres und vages Wissen, S. 291–348. München : Oldenbourg-Verlag, 2003

[Ciampi u. a. 2000] Ciampi, A. ; Diday, E. ; Lebbe, J. ; Périnel, E. ; Vignes, R.: Growing a tree classifier with imprecise data. In: *Pattern Recognition Letters* 21 (2000), Nr. 9, S. 787–803

[Diday 2008] Diday, Edwin: *Symbolic Data Analysis and the SODAS Software.* Kap. The State of the Art in Symbolic Data Analysis: Overview and Future, S. 3–41. New York, NY, USA : John Wiley & Sons, Ltd, 2008

[Foissner u. a. 1991] Foissner, W. ; Blatterer, H. ; Berger, H. ; Kohmann, F.: *Taxonomische und ökologische Revision der Ciliaten des Saprobiensystems - Band I: Cyrtophorida, Oligotrichida, Hypotrichia, Colpodea.* München : Bayerisches Landesamt für Wasserwirtschaft, 1991 (Informationsberichte des Bayer. Landesamtes für Wasserwirtschaft, 1/91)

[Hagedorn 2007] Hagedorn, G.: *Strukturierung organismischer Beschreibungsdaten - Anforderungsanalyse und Informationsmodelle.* Königin-Luise Str. 19, 14195 Berlin, Germany, Fakultät Biologie, Chemie und Geowissenschaften der Universität Bayreuth, Dissertation, Juni 2007

[Kass 1980] Kass, G. V.: An Exploratory Technique for Investigating Large Quantities of Categorical Data. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 29 (1980), Nr. 2, S. 119–127

[Miller u. a. 2009] Miller, F. P. ; Vandome, A. F. ; McBrewster, J.: *DNA Barcoding*. Alphascript Publishing, 2009

[Noirhomme-Fraiture und Brito 2011] Noirhomme-Fraiture, Monique ; Brito, Paula: Far beyond the classical data models: symbolic data analysis. In: *Statistical Analysis and Data Mining* 4 (2011), Nr. 2, S. 157–170

[Puppe 1990] Puppe, F.: *Problemlösungsmethoden in Expertensystemen*. Springer Verlag, 1990. – Studienreihe Informatik

[Puppe u. a. 2003] Puppe, F. ; Stoyan, H. ; Studer, R.: *Handbuch der Künstlichen Intelligenz*. Kap. 15. Knowledge Engineering, S. 599–642. München : Oldenbourg-Verlag, 2003

[Quinlan 1986] Quinlan, J. R.: Induction of Decision Trees. In: *Machine Learning* 1 (1986), Nr. 1, S. 81–106

[Shannon 2001] Shannon, C. E.: A mathematical theory of communication. In: *SIGMOBILE Mob. Comput. Commun. Rev.* 5 (2001), Nr. 1, S. 3–55

Glossar

Abbundanzziffer ein Maß für Individuendichte eines Taxons, Seite 12

Bestimmungsschlüssel Werkzeug oder Verfahren zur Klassifikation von biologischen Individuen, Seite 4

Evidenz ein Maß für Gewissheit, Seite 11

Klassifizierung Zuordnung eines Objektes zu einer Klasse, Seite 3

Merkmal Im taxonomischen Sinne ein Konzept für das ein Beobachtungs- oder Messverfahren festgelegt wurde, welches wiederholbare Ergebnisse eines bestimmten Typs erzeugt, Seite 4

Merkmalsausprägung Ergebnis der Anwendung eines Beobachtungs- oder Messverfahrens für ein Merkmal, Seite 4

Saprobie Maß für den Gehalt von organischen, leicht unter Sauerstoffverbrauch abbaubaren Substanzen im Wasser, Seite 53

Saprobienindex Index zur Bewertung der saprobiellen Belastung von Gewässern anhand von Indikatororganismen, Seite 1

Saprobier ein Organismus, der in Wasser lebt, in dem fäulnisfähige Stoffe enthalten sind, Seite 53

Schwellenwert Grenze, ab der ein bestimmter Wert angenommen wird, Seite 12

Sicherheitsfaktor subjektive Wahrscheinlichkeit der Gültigkeit einer Hypothese, Seite 11

Spezies, Art Klassifikationsebene der taxonomischen Systematik, Seite 3

Taxa Mehrzahl von Taxon, Seite 4

Hiermit versichere ich, dass ich die vorliegende Arbeit im Sinne der Prüfungsordnung nach §22 (4) ohne fremde Hilfe selbständig verfasst und nur die angegebenen Hilfsmittel benutzt habe.

Hamburg, 23. August 2012 Tom Klonikowski