Visualisierungsinstitut der Universität Stuttgart
Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit Nr. 22

Automatische Klassifizierung von Kontakten auf Smartphones

Huy Viet Le

Studiengang:	Softwaretechnik
Prüfer:	Prof. Dr. Albrecht Schmidt
Betreuer:	Dipl.-Inform. Niels Henze Alireza Sahami M.Sc.
begonnen am:	31. Mai 2012
beendet am:	30. November 2012
CR-Klassifikation:	H.1.2, I.5.2

Kurzfassung

Mobiltelefone ermöglichen es in der heutigen Zeit für Jedermann und jederzeit erreichbar zu sein. Dies führt dazu, dass sich in der Kontaktliste unter anderem Freunde, Familienmitglieder oder auch lose Bekanntschaften ansammeln. Dadurch wird die Kontaktliste immer unübersichtlicher und dementsprechend wird es schwieriger, einen bestimmten Kontakt ausfindig zu machen. Da aktuelle Mobiltelefone noch keine Möglichkeit bieten, Kontakte automatisch in Kontaktgruppen einzuordnen, erfordert die manuelle Einordnung viel Mühe von den Anwendern. Aus diesem Grund werden in der vorliegenden Arbeit Klassifikatoren entwickelt, welche Kontakte anhand ihrer Kommunikationsaktivität automatisch in Gruppen einordnen. Dazu werden zunächst im Rahmen einer Studie ermittelt, welche Kontaktgruppen bevorzugt von Anwendern zur Gruppierung verwendet werden. Darauf aufbauend wird anschließend ein System entwickelt, welches Kontakte inklusive ihrer Kommunikationsaktivität auf Mobiltelefonen ausliest. Die ausgelesenen Kontakte werden daraufhin manuell von Probanden in Kontaktgruppen eingeordnet. Die dadurch gesammelten Daten werden ausgewertet und letztendlich zur Entwicklung der Klassifikatoren eingesetzt. Evaluationen durch Probanden zeigen, dass die Klassifikatoren dabei bis zu 73% aller Kontakte in ihre korrekte Kontaktgruppe einordnen. Aus einem Vergleich mit sogenannten ZeroR-Klassifikatoren resultiert, dass die entwickelten Klassifikatoren bis zu 32% mehr an korrekt klassifizierten Kontakten aufweisen.

Abstract

Mobile phones enable us to be connected anywhere and anytime. The wide dissemination of mobile phones, however, also increased the number of contacts that can assemble in a user's contact list. Contact lists contain the most intimate friends, colleagues, and family members but also other persons that only have a lose connection with the user. Current systems lack an understanding of the relation between a user and her or his contacts. They require that users categorize their contacts manually. So in this thesis a classifier will be developed that automatically classifies the user's contacts using only the SMS and telecommunication history. Firstly, groups that are commonly used to categorize contacts are investigated in an online survey. On this basis a system will be established which is able to read out the contacts and activity data of every participant on a mobile phone as a part of a study. Afterwards, participants manually arrange their contacts into respective groups. The thereby collected data will be evaluated and a model will be built that classifies contacts into groups. Evaluations through probands indicate that the classifiers are able to categorize contacts correctly up to 73%. A comparison between these classifiers and so-called ZeroR classifiers demonstrates that the classifiers which were created in this thesis lead to a result better up to 32%.

Inhaltsverzeichnis

1. Einleitung **9**
1.1. Zielsetzung der Arbeit . . . 9
1.2. Aufbau der Arbeit . . . 10

2. Grundlagen **11**
2.1. Verwandte Arbeiten . . . 11
2.1.1. Wissenschaftliche Arbeiten . . . 11
2.1.2. Kommerzielle Systeme . . . 13
2.1.3. Fazit . . . 13
2.2. Technische Grundlagen . . . 14
2.2.1. Android . . . 14
2.2.2. Maschinelles Lernen . . . 15

3. Ermittlung von Kontaktgruppen **19**
3.1. Methodik . . . 19
3.2. Aufbereitung der gesammelten Antworten . . . 20
3.3. Ergebnisse . . . 21
3.3.1. Größe der Kontaktliste . . . 21
3.3.2. Nutzungshäufigkeit von SMS-Nachrichten und Telefonaten . . . 22
3.3.3. Bevorzugte Kontaktgruppen für die Gruppierung von Kontakten . . . 22
3.3.4. Gemeinsames Auftreten von Gruppen . . . 23
3.3.5. Anzahl benötigter Kontaktgruppen für die Gruppierung . . . 24
3.4. Diskussion . . . 26

4. System zur Sammlung von Kontakt- und Aktivitätsdaten **27**
4.1. Architektur . . . 27
4.2. Funktionalität . . . 28
4.3. Implementierung . . . 29
4.3.1. Funktionalitäten des Systems . . . 30
4.3.2. Format zur Speicherung der gesammelten Daten . . . 33
4.4. Zusammenfassung . . . 33

5. Sammlung und Auswertung von Kontakt- und Aktivitätsdaten **35**
5.1. Methodik . . . 35
5.1.1. Ablauf . . . 35
5.1.2. Teilnehmer . . . 36
5.2. Ergebnisse . . . 36

5.3. Diskussion 41

6. Entwicklung von Klassifikatoren **45**
6.1. Attribute 45
6.2. Entwicklung 48
6.2.1. Standard J48 Klassifikator 49
6.2.2. Multilayer J48 Klassifikator 50
6.2.3. Weitere Klassifikatoren 51
6.3. Evaluation 53
6.3.1. Vergleich mit ZeroR 53
6.3.2. Evaluation durch Probanden 54
6.4. Fazit 56

7. Fazit und Ausblick **59**
7.1. Fazit 60
7.2. Ausblick und weitere Arbeit 61

A. Appendix **63**
A.1. Beispiel einer ARFF Datei 63
A.2. Visualisierung gemeinsam vorkommender Kontaktgruppen 63
A.3. Fragebogen 65
A.4. Verwendung des Hilfsprogramms 66
A.4.1. Konvertierung ins ARFF-Format und Einführung neuer Attribute . . . 66
A.4.2. Suche nach besten Obergruppen für den Multilayer J48 Klassifikator . 67
A.5. Inhalt der CD-ROM 67

Literaturverzeichnis **69**

Abbildungsverzeichnis

2.1. Beispiel für einen Entscheidungsbaum 17
2.2. Beispiel einer Confusion Matrix 18

3.1. Nutzungshäufigkeit von Anrufen und SMS-Nachrichten bei 18-25 Jährigen . . 23
3.2. Nutzungshäufigkeit von Anrufen und SMS-Nachrichten bei 18-25 Jährigen . . 23
3.3. Visualisierung gemeinsam vorkommender Kontaktgruppen in Antworten der Teilnehmer 25

4.1. Screenshot der Android-Anwendung zur Sammlung von Kontakt- und Aktivitätsdaten 29

5.1. Verteilung der vorgegebenen Kontaktgruppen für die Gruppierung von Kontakten durch Probanden. 37
5.2. Übersicht über die Anzahl an SMS-Nachrichten und Anrufen aller Probanden. 37
5.3. Verhältnis zwischen SMS-Aktivitäten und Anruf-Aktivitäten 39
5.4. Verteilung von SMS-Aktivitäten über den Tag. 42
5.5. Verteilung von SMS-Aktivitäten über den Tag. 42
5.6. Rankinggruppen für SMS- und Anruf-Aktivitäten. 44

6.1. Erkennungsqualität des Standard J48 Klassifikators in der standard Cross-Validation und in der Cross-Validation einzelner Probanden. 49
6.2. Bestes und schlechtestes Ergebnis für die Klassifizierung eines einzelnen Folds in der Cross-Validation einzelner Probanden. 50
6.3. Hintereinanderschaltung der Klassifikatoren A, B und C zum Multilayer J48 Klassifikator. 51
6.4. Erkennungsqualität des Multilayer J48 Klassifikator in der Cross-Validation einzelner Probanden. 52
6.5. Weitere Lernmethoden und die daraus resultierenden Klassifizierungsergebnisse 52
6.6. Erkennungsqualität von ZeroR anhand der standard Cross-Validation und der Cross-Validation einzelner Probanden. 53
6.7. Confusion Matrix der Evaluierung durch Probanden. 56

Tabellenverzeichnis

3.1. Größe der Kontaktliste von Probanden (bzgl. Fragebogen) 22
3.2. Bevorzugte Kontaktgruppen für die Gruppierung von Kontakten 24
3.3. Anzahl benötigter Kontaktgruppen in Abhängigkeit zur Größe der Kontaktliste. 25

5.1. Anzahl an SMS-Aktivitäten jeweiliger Kontaktgruppen. 38
5.2. Anzahl an Anruf-Aktivitäten jeweiliger Kontaktgruppen. 38
5.3. Durchschnittliche Länge vn SMS-Nachrichten in den jeweiligen Gruppen . . . 40
5.4. Durchschnittliche Dauer von Telefonaten jeweiliger Gruppen 40

6.1. Direkt aus den gesammelten Kontaktdaten auslesbare Variablen 45
6.2. Auf die Variablen in Tabelle 6.1 anwendbaren Funktionen. 46

1. Einleitung

Die Verbreitung von Mobiltelefonen nimmt in den letzten Jahren massiv zu und hat dafür gesorgt, dass viele Menschen für Jedermann und jederzeit erreichbar sind. Insbesondere die zunehmende Verbreitung von Smartphones hat dazu geführt, dass Benutzer eine sehr große Menge an Kontakten in ihren Mobiltelefonen gespeichert haben. In der Kontaktliste finden sich die besten Freunde, Arbeitskollegen, Familienmitglieder oder aber auch Personen, zu denen der Benutzer nur eine lose Verbindung hat. Diese große Menge an Personen kann es schwierig machen eine bestimmte Person in der Kontaktliste zu finden. Aus Nutzersicht wäre es deshalb wünschenswert, wenn die Kontakte im Adressbuch in Gruppen eingeteilt wären, die das Adressbuch strukturieren und so das Finden von Personen erleichtern. Dies würde es zudem ermöglichen die eigene Kontaktierbarkeit davon abhängig zu machen in welcher Gruppe sich andere Personen befinden. Beispielsweise könnte das Telefon nur klingeln wenn eine Person aus der Gruppe Arbeitskollegen anruft und ansonsten nur durch Vibration den Anruf anzeigen. Ferner ist es im Zeitalter von Smartphones nicht unüblich private Informationen durch soziale Anwendungen mit anderen Kontakten zu teilen. Auch in diesem Fall wäre eine Gruppierung der Kontaktliste sinnvoll, um so den Überblick über die Kontakte zu verschaffen, mit denen die Informationen geteilt werden sollen.

Aus diesem Grund werden in der vorliegenden Arbeit sogenannte „Klassifikatoren“ mithilfe Techniken des maschinellen Lernens entwickelt, welche die Gruppierung der Kontakte auf Smartphones automatisch durchführen. Dabei werden zunächst Informationen zur manuellen Gruppierung von Anwendern gesammelt, welche anschließend für die Entwicklung der Klassifikatoren und zur Lösung dieser Aufgabe eingesetzt werden.

1.1. Zielsetzung der Arbeit

Ziel dieser Arbeit ist die Entwicklung eines Klassifikators, welcher Kontakte, die sich im Adressbuch von Smartphones befinden, automatisch in Kontaktgruppen einordnen kann. Dabei werden auf die von Smartphones bereits heute gesammelten Daten zurückgegriffen. Insbesondere relevant sind beispielsweise die getätigten Anrufe, die versendeten SMS sowie die Uhrzeiten, an denen das Smartphone zur Kommunikation verwendet wurde. Für die Realisierung werden zwei Studien zur Sammlung empirischer Daten durchgeführt. In der ersten Studie wird ermittelt, welche Kontaktgruppen bevorzugt für die Gruppierung von Kontakten verwendet werden. Darauf aufbauend werden in der zweiten Studie Daten gesammelt, zu denen Kontaktinformationen mit ihren Aktivitäten sowie die manuelle Gruppierung durch Probanden gehören. In diesem Schritt wird eine mobile Anwendung erstellt, welche die genannten Daten aus Smartphones exportiert und diese für die weitere

Arbeit verfügbar macht. Nach der Sammlung der Daten werden diese ausgewertet und letztendlich für die Entwicklung der Klassifikatoren verwendet. Anschließend folgt eine Evaluation der entwickelten Klassifikatoren.

1.2. Aufbau der Arbeit

Die vorliegende Arbeit beginnt in Kapitel 2 mit einer Recherche über verwandte Arbeiten sowie einer Einführung in die technischen Grundlagen. Anschließend wird in Kapitel 3 im Rahmen einer Studie ermittelt, welche Kontaktgruppen bevorzugt zur Gruppierung der Kontaktliste verwendet werden. Zudem wird in der Studie das Nutzungsverhalten der Probanden analysiert. Darauf aufbauend wird im nachfolgenden Kapitel 4 die nächste Studie vorbereitet, indem ein System zur Sammlung von empirischen Daten entwickelt wird. Dieses System wird in Kapitel 5 eingesetzt, um Kontakt- und Aktivitätsdaten von Probanden zu sammeln, die anschließend ausgewertet werden. Nach der Auswertung beginnt die Entwicklung der Klassifikatoren, welche im Anschluss evaluiert werden. Die Arbeit endet im letzten Kapitel mit einer Zusammenfassung, dem Fazit sowie einem Ausblick auf mögliche Weiterentwicklungen oder Einsätze der Klassifikatoren.

2. Grundlagen

In diesem Kapitel werden Einblicke über verwandte Arbeiten im Bereich sowie kommerzielle Systeme gegeben. Technische Grundlagen, die für das Verständnis dieser Arbeit elementar sind, werden hier kurz zusammengefasst.

2.1. Verwandte Arbeiten

Es folgt in den folgenden Abschnitten eine Vorstellung über bereits existierende wissenschaftliche Arbeiten sowie kommerzielle Systeme. Anschließend folgt ein Fazit über die vorgestellten Arbeiten sowie die Unterscheidung zur vorliegenden Arbeit.

2.1.1. Wissenschaftliche Arbeiten

Es existieren bereits verschiedene Arbeiten, welche sich mit der Organisation von Kontakten auf Smartphones beschäftigen. So präsentieren Komninos und Liarokapis vier Faktoren in ihrer Arbeit, welche die Suche nach einem Kontakt im Telefonbuch des Smartphones beschleunigen sollen. [KL09] Zu diesen Faktoren gehören die Häufigkeit und Aktualität der Kommunikation mit einem Kontakt, der aktuelle Ort und die Uhrzeit in Bezug auf den Kontext, anstehende Aufgaben und Aktivitäten sowie die persönliche Präferenz des Anwenders. Ebenfalls mit der Verbesserung des Telefonbuchs beschäftigten sich Oulasvirta et al. in ihrer Arbeit *ContextContacts* [ORT05], in der das Telefonbuch um kontextbezogene Angaben der Kontakte ergänzt werden. Um wichtige Kontakte im Telefonbuch hervorzuheben, stellen Stefanis et al. in [SPKG12] eine Methode zur Erkennung von wichtigen Kontakten auf Smartphones vor. Zwei Faktoren spielen dabei eine große Rolle: Einerseits die Häufigkeit und Aktualität der Kommunikation mit einem Kontakt sowie die persönliche Präferenz des Benutzers, zu der unter anderem das manuelle Favorisieren zählt. Durch Studien kommen Stefanis et al. unter anderem zu einem Ergebnis, das dem oftmals zitierten Pareto Principle (80/20 Regel) entspricht: während mit maximal 20% aller Kontakte mindestens wöchentlich kommuniziert und somit als wichtig empfunden werden, werden andere Kontakte nie oder nur sehr selten kontaktiert.

Mit *Friendlee: A Mobile Application for Your Social Life* [ASL+09] präsentieren Ankolekar et al. eine Anwendung, welche mithilfe der bereits vorhandenen SMS- und Anrufhistorie ein soziales Netzwerk bildet. Pietiläinen et al. stellen mit *MobiClique: Middleware for Mobile Socail Networking* [POL+09] eine weitere Arbeit vor, die sich mit der Erstellung eines sozialen Netzwerks auf Mobiltelefonen beschäftigt. Anders als in [ASL+09] wird das soziale Netzwerk in

MobiClique über eine Bluetooth-Verbindung aufgebaut. Ebenfalls mit sozialen Netzwerken beschäftigt sich die Arbeit *Predicting Tie Strength With Social Media* [GK09]. Dabei stellen Gilbert et al. einen Algorithmus zur Berechnung der Stärke einer Beziehung zwischen Personen in sozialen Netzwerken im Internet vor. Als Einheit wurde die sogenannte *tie strength* gewählt, welches auf Granovetters Arbeit [Gra73] basiert. In die Berechnung fließen Faktoren wie Kontaktintensität, gemeinsame Interessen, gemeinsames Auftreten auf Fotos, soziale Distanz (Altersunterschied, Unterschied in der Bildung oder politische Gemeinsamkeiten) oder Verwendung von bestimmten Wörtern (mithilfe von LIWC[1]) ein. Nach Angaben von Gilbert et al. erreicht das entwickelte Modell eine Genauigkeit von bis zu 85% bei der Berechnung der *tie strength*.

Anstatt Kontakte selbstständig zu organisieren, stellen Amershi et al. in [AFW12] eine Anwendung vor, die Vorschläge während der manuellen Gruppierung der Kontaktliste in sozialen Netzwerken liefert. Dazu wurde ein System entwickelt, das Facebook Nutzern anhand von Profilangaben und Häufigkeiten der Kommunikation ähnliche Kontakte zu den bereits gruppierten Kontakten vorschlägt.

Mit der Organisation von E-Mail Kontakten beschäftigten sich MacLean et al. in ihrer Arbeit *Groups Without Tears: Mining Social Topologies from Email* [MHT+11], bei der eine Methode zur Einordnung von E-Mail Kontakten in sogenannte social topologies vorgestellt wird. Social topologies sind Mengen von potentiell überlappenden und geschachtelten sozialen Gruppen. Dazu wurden die gemeinsamen Empfänger betrachtet und daraus *Social Molecules* hergeleitet. Social Molecules stellen eine Gruppe von Kontakten dar, die bei den Empfängern stets zusammen auftreten. Diese Arbeit wurde anschließend in [PMT+11] fortgesetzt, woraus die Facebook Anwendung GroupGenie [2] entstand. Diese Facebook Anwendung erlaubt es anhand von getaggten Fotos und gemeinsamen Freunden eine Social Topology aufzustellen. Ebenfalls mit der Gruppierung von E-Mail Kontakten beschäftigten sich Roth et al. in [RBDD+10]. Es wurde ein Algorithmus veröffentlicht, der bei einer manuellen Gruppierung von E-Mail Kontakten durch Benutzer anhand von Vorschlägen helfen soll. Dazu wurde der *implicit social graph* eingeführt, welcher durch Interaktionen von Benutzern mit ihren Kontakten oder Gruppen von Kontakten aufgebaut wird. Dieser unterscheidet sich von *explicit social graphs* (bspw. OpenGraph von Facebook [3]) dadurch, dass Personen nicht *explizit* durch den Benutzer zur Kontaktliste hinzugefügt werden müssen. Das Ergebnis dieser Arbeit wurde unter anderem in den GMail Labs Features *Don't forget Bob!* und *Get the wrong Bob?* [4] eingesetzt. Diese Features schlagen bei einer E-Mail Gruppenkonversation Empfänger vor oder weisen auf einen möglicherweise falschen Empfänger hin.

[1] *LIWC - Linguistic Inquiry and Word Count*: http://www.liwc.net/ *(zuletzt geprüft: 25.11.2012)*
[2] *GroupGenie*: http://mobisocial.stanford.edu/groupgenie/ *(zuletzt geprüft: 25.11.2012)*
[3] *Facebook OpenGraph*: https://developers.facebook.com/docs/opengraph/ *(zuletzt geprüft: 25.11.2012)*
[4] *GMail Labs Features*: http://gmailblog.blogspot.de/2011/04/dont-forget-bob-and-got-wrong-bob.html *(zuletzt geprüft: 25.11.2012)*

2.1.2. Kommerzielle Systeme

Zu den kommerziellen Systemen gehört unter anderem die Anwendung *Katango* [5], welches Facebook-Kontakte anhand einer Menge von Faktoren in Gruppen einordnet. Zu diesen Faktoren gehören unter anderem Verbindungen im Kontaktgraphen, Wohnort sowie Interaktionen mit Kontakten auf Facebook. [6] Im November 2011 kaufte Google diese Anwendung, um so die Circles-Funktionen von Google+ zu erweitern. [7] Google+ [8] ist ein von Google entwickeltes soziales Netzwerk, dessen Hauptmerkmal die sogenannten Circles sind. Dies sind im Grunde Kontaktlisten, in die der Benutzer seine Kontakte einordnen muss, um Beiträge mit diesen zu teilen.

Als ein Feature von Facebook erstellt Facebook Smartlists automatisch Gruppen mithilfe von Profilangaben. Dabei werden Arbeitskollegen, Schul-/Unifreunde sowie Freunde aus einem nahen Ort anhand von Profilangaben erkannt und in die Gruppen eingeordnet. Die dadurch erstellen Gruppen sind nicht disjunkt, was heißt, dass ein Kontakt in mehreren Gruppen vorkommen kann. [9]

2.1.3. Fazit

Im Gegensatz zur vorliegenden Arbeit unterscheiden sich die verwandten Arbeiten und kommerzielle Systeme dadurch, dass sie Kontakte in sozialen Netzwerken oder im E-Mail-Verkehr, bei dem mehrere Empfänger üblich sind, kategorisieren. Da auf Mobiltelefonen weder Informationen zu gemeinsamen Freunden oder Interessen noch gemeinsame Empfänger von SMS-Nachrichten oder Gesprächspartner einer Gruppenkonversation verfügbar sind, können die vorgestellten Ansätze nicht zur Lösung dieser Arbeit eingesetzt werden.

Die Arbeiten, die sich mit der Verbesserung des Telefonbuchs beschäftigen, präsentierten hauptsächlich Algorithmen zur Erkennung von wichtigen Kontakten oder ergänzten das vorhandene Telefonbuch um kontextsensitive Informationen. Keine der Arbeiten jedoch beschäftigte sich explizit mit der Gruppierung von Kontakten in vorgegebene Kontaktgruppen. Zudem eignen sich auch die Sortierungen beziehungsweise Gruppierungen, wie sie bereits vorgeschlagen wurden, nicht für das schnelle Finden eines Kontakts bezüglich eines Kontext im Alltag. So wurde hauptsächlich nach Stärke einer Beziehung (*tie strength*) oder nach Wichtigkeit sortiert. Möchte der Benutzer jedoch möglichst schnell einen Arbeitskollegen oder ein Familienmitglied anrufen, mit dem er sonst mittelmäßig viel telefoniert oder SMS austauscht, so spielt die Wichtigkeit oder Stärke einer Beziehung bei der Suche nach dem Kontakt nahezu keine Rolle mehr.

[5] *Katango*: http://www.katango.com/ *(zuletzt geprüft: 25.11.2012)*
[6] *Katango explained*: http://www.building43.com/videos/2011/07/12/katango-organizing-your-social-network/ *(zuletzt geprüft: 25.11.2012)*
[7] *Google+ Circles*: http://support.google.com/plus/bin/answer.py?hl=de&answer=1047805 *(zuletzt geprüft: 25.11.2012)*
[8] *Google+*: http://plus.google.com/ *(zuletzt geprüft: 25.11.2012)*
[9] *Facebook Smartlists*: https://blog.facebook.com/blog.php?post=10150278932602131 *(zuletzt geprüft: 25.11.2012)*

2.2. Technische Grundlagen

In diesem Abschnitt sollen die für diese Arbeit notwendigen Grundlagen kurz und knapp erläutert werden. Dazu gehört eine Einführung in Android, eine Erläuterung elementarer Werkzeuge für die Entwicklung von Android-Anwendungen, sowie eine Einführung in das maschinelle Lernen.

2.2.1. Android

Android ist ein quelloffenes Betriebssystem für Smartphones, Tablets und Netbooks. Es wird von der Open Handset Alliance, einen Konsortium, das Google mit 33 Partnern gegründete, entwickelt. Mit einem weltweiten Marktanteil von 68,1% im zweiten Quartal 2012 [10] gehört es zu den meistverwendeten Betriebssystemen für Smartphones.

Android basiert auf dem Linux Kernel, das für die Speicher- und Prozessverwaltung zuständig ist. Die Dalvik Virtual Machine, eine für Android entwickelte Java Virtual Machine, ermöglicht es in Java geschriebene Applikationen (kurz Apps) auszuführen. Android Apps werden größtenteils durch einen Touchscreen sowie einigen Hardwaretasten[11] bedient.

Die meisten Anwendungen für Android werden im *Google Play Store* [12] (früher Google Market) angeboten. Möchte der Entwickler seine Anwendung nicht im Google Play Store veröffentlichen, so kann die Anwendung mithilfe einer APK-Datei [13] auf einem Gerät installiert werden.

Entwicklung und Aufbau einer Anwendung

Im Folgenden werden Werkzeuge zur Entwicklung sowie einige Komponenten von Android-Anwendungen vorgestellt, die für das Verständnis der weiteren Arbeit notwendig sind.

Android-Anwendungen werden in der Programmiersprache Java mithilfe des Android-SDKs [14] entwickelt. Das Android SDK stellt dabei nötige Bibliotheken zur Verfügung und unterstützt den Entwickler durch Debugging- und Testwerkzeuge. In der vorliegenden Arbeit wurde die Entwicklungsumgebung Eclipse [15] in der Version 4.2 Juno mit Einbindung des Android-SDKs verwendet. Aktuell ist das Android-SDK in der Version r21 verfügbar.

[10] *Smartphone-Weltmarkt nach Systemen*: http://www.heise.de/newsticker/meldung/Marktforscher-Ueber-100-Millionen-Androiden-ausgeliefert-1659638.html *(zuletzt geprüft: 25.11.2012)*
[11] Bei einigen Modellen wurden diese Hardwaretasten durch das Betriebssystem auf dem Bildschirm simuliert.
[12] *Google Play Store*: https://play.google.com/store *(zuletzt geprüft: 25.11.2012)*
[13] APK (**A**ndroid **P**ac**k**age) ist ein Dateiformat, mit der Android-Anwendungen installiert werden können
[14] *Android SDK*: http://developer.android.com/sdk/index.html *(zuletzt geprüft: 25.11.2012)*
[15] *Eclipse SDK*: http://www.eclipse.org/ *(zuletzt geprüft: 25.11.2012)*

Die Oberflächen von Android Anwendungen werden durch sogenannte *Activities* [16] realisiert. Diese geben dem Programmierer die Möglichkeit verschiedene Steuerelemente darauf zu platzieren. Eine Anwendung kann mehrere Activities besitzen, zwischen denen gewechselt werden kann. Im Lebenszyklus einer Anwendung kann jedoch immer nur eine Activity zur selben Zeit aktiv sein.

Mithilfe eines Intents [17] kann zwischen verschiedenen Activities gewechselt werden. Es besteht dabei die Möglichkeit Daten zwischen den Activities auszutauschen, indem diese an den Intent angehängt werden. Um Daten über Anwendungsgrenzen auszutauschen, werden Content Provider [18] verwendet. Beispielsweise werden Content Provider in der weiteren Arbeit verwendet, um SMS-Nachrichten oder den Anrufverlauf auf einem Android System auszulesen. Eine Anwendung kann jederzeit auf einen vorhandenen ContentProvider über eine URI (Uniform Resource Identifier) der Form `content://<Basisadresse>/<Pfad>[/<id>]` zugreifen.

Für den Zugriff auf bestimmte Content Provider (beispielsweise SMS-Speicher und Anruf-historie) werden Berechtigungen benötigt. Diese Berechtigungen werden bei Installation einer Anwendung angezeigt und müssen anschließend vom Benutzer bestätigt werden. Programmierer müssen diese in der Manifest-Datei (AndroidManifest.xml [19]) eintragen. [20]

2.2.2. Maschinelles Lernen

In diesem Abschnitt werden Begriffe aus dem Bereich *Maschinelles Lernen* erläutert, die für das Verständnis der weiteren Arbeit relevant sind. Da eine Software für die Auswertung und Entwicklung des Klassifikators verwendet wird, bei der die im Folgenden erläuterten Entscheidungsbäume bereits implementiert und zur Anwendung bereit stehen, wird hier auf die Erläuterung der Details zur Implementierung verzichtet.

Klassifikatoren

Mathematisch gesehen ist ein Klassifikator eine Funktion $f(o) \rightarrow k$, die ein Objekt o anhand seiner Attribute in eine Kategorie k einordnet. Zusätzlich ist in der vorliegenden Arbeit mit einem *Klassifikator* ein Algorithmus gemeint, dem eine Lernphase ("Training") vorausging.

[16] *Android SDK (Activity)*: http://developer.android.com/reference/android/app/Activity.html *(zuletzt geprüft: 25.11.2012)*
[17] *Android SDK (Intent)*: http://developer.android.com/guide/components/intents-filters.html *(zuletzt geprüft: 25.11.2012)*
[18] *Android SDK (Content Provider)*: http://developer.android.com/guide/tographics/providers/content-provider-basics.html *(zuletzt geprüft: 25.11.2012)*
[19] *Android SDK (AndroidManifest.xml)*: http://developer.android.com/guide/topics/manifest/manifest-intro.html *(zuletzt geprüft: 25.11.2012)*
[20] *Android SDK (Manifest Permissions)*: http://developer.android.com/reference/android/Manifest.permission.html *(zuletzt geprüft: 25.11.2012)*

Dabei werden dem Klassifikator Trainingsdaten zur Verfügung gestellt, mit dem durch verschiedene Lernmethoden Merkmale ausfindig gemacht werden, um Objekte einer Kategorie einordnen zu können. [WF05, S. 43]

Lernmethoden

Klassifikatoren können auf unterschiedliche Weise von ihren Trainingsdaten lernen. Die Art und Weise, wie ein Klassifikator lernt, wird Lernmethode genannt. Die Wahl der Lernmethode ist abhängig von der Art der Trainingsdaten.

Alle Lernmethoden vorzustellen wäre im Rahmen dieser Arbeit zu ausführlich. Daher werden hier ausschließlich Entscheidungsbäume als Lernmethode erklärt werden, da diese in den weiteren Kapiteln noch zum Einsatz kommen werden. Einführungen in weitere Lernmethoden werden in [Run10, S. 85ff] gegeben.

Entscheidungsbäume

Mithilfe von Entscheidungsbäumen kann ein Objekt anhand einer vorgegebenen Baumstruktur einer Kategorie zugeordnet werden. Um ein Objekt zu kategorisieren, wird, beginnend von der Wurzel, ein Pfad im Baum verfolgt. Durch Entscheidungsfragen an den inneren Knoten des Baumes wird der Pfad vorgegeben. Schlussendlich geben Blätter die Kategorien an, in die ein Objekt eingeordnet wurde. [WF05, S. 62] (in diesem Falle die Kategorien *Ja* und *Nein*)

Für die Generierung von Entscheidungsbäumen wird in der vorliegenden Arbeit der in WEKA implementierte J48-Algorithmus [Qui93] verwendet. Der J48-Algorithmus ist eine in Java implementierte, sowie verbesserte Version des C4.5 Algorithmus'. [WF05, S. 373]

In Abbildung 2.1 ist ein Entscheidungsbaum abgebildet, welcher alle Instanzen aus der ARFF-Datei in Appendix A.1 kategorisiert. In diesem aus [WF05, S. 101] ins Deutsche übersetzte Beispiel soll anhand des Wetters, welches aus einer Kombination mehrerer Attribute besteht, bestimmt werden, ob bei diesem Wetter Tennis gespielt werden kann oder nicht. Dabei wird zunächst das Attribut *Aussicht* betrachtet. Ist die Aussicht bewölkt, so liefert der Entscheidungsbaum die Antwort (Kategorie) *Ja*. Falls die Aussicht sonnig oder regnerisch ist, werden anschließend die Attribute *Feuchtigkeit* (wenn sonnig) oder *Windigkeit* (wenn regnerisch) abgefragt. Je nach *Feuchtigkeit* beziehungsweise *Windigkeit* wird anschließend entschieden, ob Tennis gespielt werden kann oder nicht.

Beurteilung von Klassifikatoren

Die Güte (Qualität) eines entwickelten Klassifikator kann durch die Werte *Precision*, *Recall* und den *F-Measure* angegeben werden. Diese Werte setzen sich aus True Positives (TP), False Positives (FP) und False Negatives (FN) zusammen. *TP* gibt die Anzahl an Objekten an, die korrekterweise einer Klasse zugeordnet worden sind. *FP* sind die Objekte, die

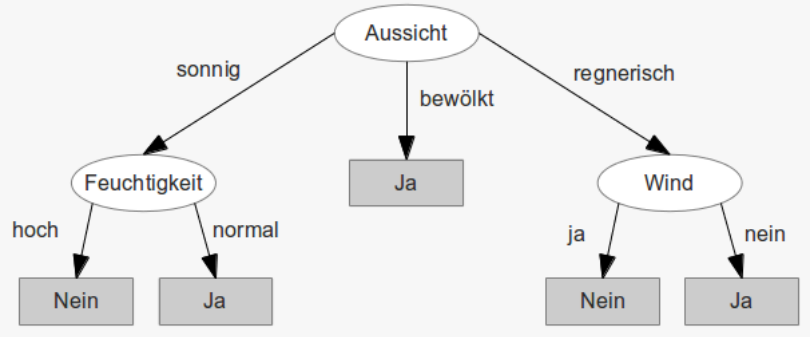


Abbildung 2.1.: Beispiel für einen Entscheidungsbaum. In diesem Beispiel wird durch die Attribute Aussicht, Feuchtigkeit und Wind bestimmt, ob Tennis gespielt wird oder nicht.

fälschlicherweise einer Klasse zugeordnet worden sind, während *FN* die Objekte zeigt, die fälschlicherweise nicht in eine Klasse eingeordnet wurden. [MRS08]

In der vorliegenden Arbeit dient der F-Measure-Wert als Metrik für die Güte eines Klassifikators. Folgende Tabelle gibt an, wie die genannten Werte berechnet werden:

	Precision	**Recall**	**F-Measure**
Beschreibung	Gibt den Anteil an korrekt klassifizierten Objekten zur Klasse k an.	Gibt den Anteil der zur Klasse k gehörenden Objekte an, die vom Klassifikator auch wirklich als k klassifiziert wurden.	Kombination von Precision und Recall
Berechnung	$\frac{TP_k}{TP_k+FP_k}$	$\frac{TP_k}{TP_k+FN_k}$	$\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$

Cross-Validation

Die Cross-Validation (zu Deutsch: Kreuzvalidierung) ist ein Testverfahren, bei dem die Güte eines Klassifikators bestimmt wird. Dabei werden alle Objekte, die für die Entwicklung des Klassifikators zur Verfügung stehen, zunächst in eine zufällige Reihenfolge gebracht. Anschließend wird in n Iterationen jeweils ein Klassifikator entwickelt, wobei n-1 Sets für das Training und das übrig gebliebene Set für den Test des Klassifikators verwendet werden. Von allen Testergebnissen werden anschließend die Durchschnittswerte berechnet, die schlussendlich die Güte (s. Beurteilung von Klassifikatoren) des Modells angeben. [Koh95]

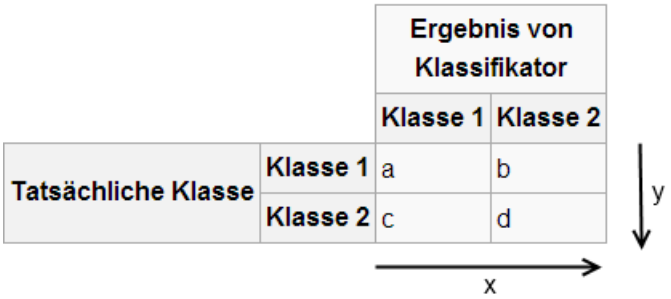


Abbildung 2.2.: Beispiel für eine Confusion Matrix. In diesem Bespiel gibt die Zelle a die Anzahl an True Positives an, während Zelle b und c die False Positives und die False Negatives angeben.

Confusion Matrix

Zum Ergebnis einer Cross-Validation gehört unter anderem die Confusion Matrix [VRRK11] (zu Deutsch: Konfusionsmatrix), die mit einem tabellenorientierten Layout die Klassen sowie die dazugehörigen Klassifizierungen der Objekte durch einen Klassifikator visualisiert. Dabei geben die Spalten x an, in welche Klasse ein Objekt klassifiziert wurde, während die Zeilen y angeben, zu welcher Klasse ein Objekt tatsächlich gehört. (s. Abbildung 2.2)

Alle Zellen, für die $x = y$ gilt, geben somit die Anzahl an korrekt klassifizierten Objekten (True Positives) der Klasse x (und y) an. Im Beispiel sind dies die Zellen a und d. Die Zahlen in Zelle (x, y), für die $x \neq y$ gilt, geben die Anzahl an *False Positives* an, welche in Klasse x statt in ihre korrekte Klasse y klassifiziert wurden (Zelle b und c). Gleichzeitig geben die genannten Zellen die Anzahl an *False Negatives* an, welche fälschlicherweise *nicht* in die Klasse y eingeordnet wurden.

WEKA

WEKA [21] [WFT+99] steht für "Waikato Environment for Knowledge Analysis" und ist eine in Java implementierte Software, die verschiedene Techniken aus den Bereichen Maschinelles Lernen und Data-Mining bereitstellt. Dazu gehören zahlreiche Lernmethoden, Filter um eingelesene Daten vor der Trainingsphase vorzubereiten sowie Funktionalitäten zum Testen der entwickelten Klassifikatoren. (s. Cross-Validation) Trainingsdaten für die Klassifikatoren werden im ARFF-Format eingelesen. Ein Beispiel für eine ARFF-Datei ist in Appendix A.1 abgedruckt. In der vorliegenden Arbeit wird WEKA in der Version 3.7.7 verwendet.

[21] *WEKA*: http://www.cs.waikato.ac.nz/ml/weka/ *(zuletzt geprüft: 25.11.2012)*

3. Ermittlung von Kontaktgruppen

In diesem Kapitel wird ermittelt, welche Kontaktgruppen von Probanden bevorzugt zur Gruppierung ihrer Kontakte auf Smartphones verwendet werden. Dadurch kann in der weiteren Arbeit auf vorgegebene Kontaktgruppen zurückgegriffen werden, welche nicht vom Autor dieser Arbeit, sondern von Probanden vorgeschlagen wurden. Neben den bevorzugten Kontaktgruppen wird auch das Nutzungsverhalten der Smartphone-Besitzer analysiert und die durchschnittlich benötigte Anzahl an Kontaktgruppen berechnet. Das daraus gewonnene Wissen wird als Grundlage für die weitere Arbeit eingesetzt.

Dazu wird eine Umfrage durchgeführt, in der die bevorzugten Kontaktgruppen, sowie weitere Angaben, wie die Häufigkeit der Verwendung von SMS und Anrufen oder die Größe der Kontaktliste, ermittelt werden. Es folgt anschließend die Auswertung des Fragebogens und eine Diskussion.

3.1. Methodik

Die Umfrage wurde mithilfe von *Google Forms* [1] realisiert. Dieser Service ermöglicht unter anderem eine Verteilung des Fragebogens über das Internet, sowie das Herunterladen der Antworten in Form einer CSV-Datei [2]. Nach Fertigstellung des Fragebogens wurde dieser über soziale Netzwerke (u.a. Facebook oder Google+), Mailinglisten sowie über mündliche Wege verteilt und war eine Woche lang vom 12. Juni 2012 bis zum 19. Juni 2012 verfügbar.

Der Fragebogen mit Nummerierung der Fragen ist in Appendix A.3 abgedruckt. Dieser beinhaltet neben demographischen Fragen zu Alter und Geschlecht eine Frage zur Größe der Kontaktliste (Frage 1), die Häufigkeit der Kommunikation durch Anrufe (Frage 2) und SMS (Frage 3) sowie die bevorzugten Kontaktgruppen für die Gruppierung der Kontaktliste (Frage 4).

An dieser Umfrage nahmen insgesamt 82 Personen teil. Von allen Teilnehmern waren 44% (36 Teilnehmer) männlich und 56% (46 Teilnehmer) weiblich. Die Altersgruppe 18-25 Jahre war mit 56 Teilnehmern zu 65% vertreten, während die Altersgruppe 26-50 Jahre mit 29 Teilnehmern zu 35% vertreten wurde. Andere zur Verfügung stehende Altersgruppen (jünger als 14, 14-17 und älter als 50) wurden nicht gewählt.

[1] *Google Forms*: http://www.google.com/drive/start/apps.html *(zuletzt geprüft: 25.11.2012)*

[2] Comma-separated values. Alle Werte werden dabei mit Kommata voneinander getrennt. Für eine genauere Erläuterung siehe: https://www-cgi.tu-darmstadt.de/wiki/www/doku.php?id=csv. *(zuletzt geprüft: 25.11.2012)*

3.2. Aufbereitung der gesammelten Antworten

Da die Antworten zur Frage nach den bevorzugten Kontaktgruppen als Freitext vorlagen, mussten diese zunächst aufbereitet werden. Dazu gehören neben einigen Formkorrekturen auch das Entfernen nicht auswertbarer Antworten sowie eine Vereinheitlichung der Gruppenbezeichner. Im Folgenden werden diese Schritte genauer erläutert. Zudem werden die Namen der Kontaktgruppen ab jetzt mit *Gruppenbezeichner* benannt.

Entfernung nicht auswertbarer Antworten

Zwei Teilnehmer gaben an, keine Kontaktgruppen für ihre Kontakte zu verwenden und gaben deshalb folgende Antworten:

- "Keine :)"
- „I would not organize them in this way. An alphabetical list (or an ordered list based on recency of communication) is enough for me." (*Freie Übersetzung*: "Ich würde meine Kontakte nicht auf diese Weise organisieren. Eine alphabetisch geordnete Liste (oder eine, die nach den zuletzt verwendeten Kontakten sortiert ist) ist für mich ausreichend.")

Da diese Antworten keine Gruppenbezeichner sind, wurden diese für die weitere Auswertung nicht betrachtet.

Vereinheitlichung von Gruppennamen

Es wurden Gruppenbezeichner angegeben, welche bedeutungsgleich zu anderen bereits genannten Gruppenbezeichner sind. In diesem Abschnitt sollen alle Gruppenbezeichner vereinheitlicht werden, indem Synonyme durch einen eindeutigen Begriff ersetzt werden. Zudem wurden von 16 Probanden englische Gruppenbezeichner angegeben, welche ins Deutsche übersetzt wurden.

Es folgt eine Aufzählung aller eindeutigen Gruppenbezeichner. Unterhalb der eindeutigen Gruppenbezeichner werden die Synonyme oder die englische Bezeichner aufgeführt, welche durch den eindeutigen Gruppenbezeichner ersetzt werden.

- *Eindeutiger Gruppenbezeichner*: Enge Freunde
  Synonyme: Favoriten, VIP, Lieblingsmenschen, Friends i meet regularly, Mädels[3]

[3] Wenn dieser Gruppenbezeichner von weiblichen Teilnehmern genannt wurde und diese ebenfalls die Gruppe „Freunde" genannt haben

- *Arbeit*
  Work, Arbeitskollegen, Kollegen, Colleagues, Firma, Geschäftlich, Job, Büro, geschäftliche Kontakte, Geschäftskollegen, Geschäftskolegen, Kunden, Arbeite Kontakte, Business
- *Andere*
  Sonstiges, Others, Other, Diverses, Misc, Etc
- *Uni (+ Schule und Ausbildung)*
  Kommilitonen, Schule
- *Hobby*
  Training, Sport, Reiten, Hobbygruppe, Sport-/Musikverein, Verein, Reitsport, Stallgemeinschaft
- *Services*
  Dienste, „Servicenummern (ADAC, Hausmeister, Taxi, Andere)“, Behörden, „Offizielle (Ärzte..)“

3.3. Ergebnisse

Im Folgenden werden alle Antworten ausgewertet und aufgelistet. Zunächst werden die Größen der Kontaktlisten, das Nutzungsverhalten und die meistverwendeten Gruppenbezeichner ausgewertet. Anschließend wird das gemeinsame Auftreten von Gruppenbezeichnern sowie die Abhängigkeit zwischen Gruppenanzahl und Kontaktanzahl behandelt.

3.3.1. Größe der Kontaktliste

Bei der Frage nach der Größe ihrer Kontaktliste gaben 20 Probanden an, weniger als 50 Kontakte in dieser gespeichert zu haben, während dagegen 35,4% der Teilnehmer zwischen 50 und 100 Kontakte besaßen. Bei 26,8% sind eine Anzahl zwischen 200 und 400 Kontakten vorhanden, allerdings konnte dies nur noch von 8 Probanden mit 200-400 Kontakten und lediglich 3 Probanden mit über 400 Kontakten gesteigert werden. Diese Werte sind zusammengefasst in Tabelle 3.1 aufgelistet.

Größe der Kontaktliste	Probanden
weniger als 50 Kontakte	20 (24,4%)
zwischen 50 und 100 Kontakte	29 (35,4%)
zwischen 100 und 200 Kontakte	22 (26,8%)
zwischen 200 und 400 Kontakte	8 (9,8%)
mehr als 400 Kontakte	3 (3,6%)

Tabelle 3.1.: Antworten von Probanden zur Frage nach der Größe ihrer Kontaktliste (Auswertung von Frage 1)

3.3.2. Nutzungshäufigkeit von SMS-Nachrichten und Telefonaten

Das Nutzungshäufigkeit von Anrufen und SMS-Nachrichten wurde in Abbildung 3.1 für die beiden Altersgruppen 18-25 Jahre und 26-50 Jahre ausgewertet. Zu erkennen ist, dass 51% der 18-25-jährigen täglich mehrere SMS-Nachrichten austauschen, wobei insgesamt 77% der Altersgruppe täglich mindestens eine SMS versenden oder empfangen. 15% der Gruppe versenden hingegen nur wöchentlich SMS-Nachrichten. Lediglich 6% gaben an, dass sie monatlich Gebrauch von SMS-Nachrichten machen. Übrig bleiben 2%, welche komplett auf die Nutzung von SMS-Nachrichten verzichten.

Im Vergleich zu den 51% bei den SMS-Nachrichten tätigen nur 23% der 18-25-jährigen mehrere Anrufe täglich. 34% telefonieren einmal täglich, während ein anderer Anteil von 34% wöchentlich Telefonate führt. Insgesamt 10% gaben an, dass sie ihr Mobiltelefon monatlich oder nie für Anrufe verwenden.

In der Altersgruppe 26-50 besteht ein ähnliches Verhalten gegenüber Telefonaten mit dem Mobiltelefon. 21% der Befragten machten die Angabe, mehrmals am Tag Anrufe zu tätigen. Einmal täglich telefonieren 31% dieser Altersgruppe, wohingegen 34% einmal wöchentlich mit dem Mobiltelefon anrufen oder angerufen werden. 13% verwenden ihr Mobiltelefon entweder monatlich oder nie für das Telefonieren.

Anders als bei der Altersgruppe 18-25 gibt es bei den 26-50-jährigen keinen Vorzug für SMS-Nachrichten. Wie bei den Anrufen gaben lediglich 21% an, mehrmals täglich SMS-Nachrichten zu versenden. Weitere 31% nahmen einmal täglich die Option in Anspruch SMS-Nachrichten zu versenden. 38% gaben an nur wöchentlich Gebrauch von SMS-Nachrichten zu machen, während insgesamt 20% dies monatlich bis hin zu nie tun.

3.3.3. Bevorzugte Kontaktgruppen für die Gruppierung von Kontakten

Es wurde ausgewertet, welche Kontaktgruppen am häufigsten für die Gruppierung der Kontakte angegeben wurden (Frage 4). Wie in Tabelle 3.2 dargestellt, verwenden 79,3% aller Teilnehmer die Kontaktgruppe *Familie* für ihre Gruppierung. Dahinter folgt die Kontaktgruppe *Freunde* mit 70.7%. Die Kontaktgruppe *Arbeit* wurde von 46 Teilnehmern angegeben, was

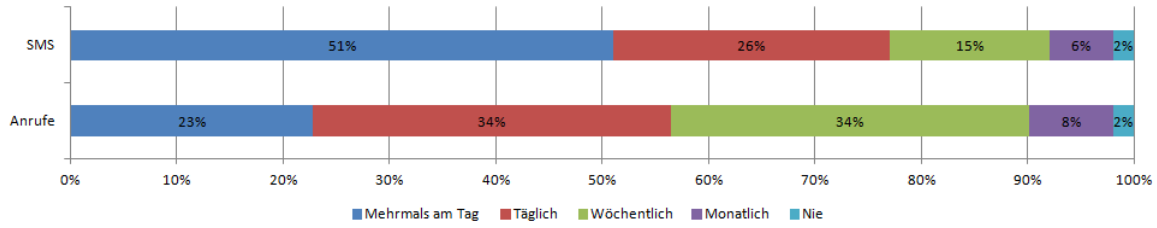


Abbildung 3.1.: Antworten zu den Fragen nach der Nutzungshäufigkeit von SMS-Nachrichten und Telefonaten bei 18-25-jährigen (Auswertung von Frage 2 und 3)

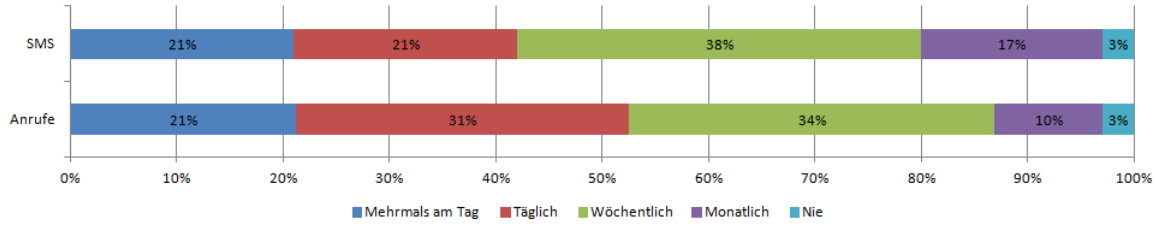


Abbildung 3.2.: Antworten zu den Fragen nach der Nutzungshäufigkeit von SMS-Nachrichten und Telefonaten bei 26-50-jährigen (Auswertung von Frage 2 und 3)

einen Anteil von 56.1% ausmacht. Nach den 3 meistverwendeten Gruppen, welche alle von mindestens der Hälfte aller Teilnehmer genannt wurden, folgt die Gruppe *Uni* mit 32.9% sowie die Kontaktgruppen *Andere* und *Bekannte* mit jeweils 22%. Nach den Kontaktgruppen *Hobby* und *Enge Freunde*, welche jeweils 11.2% und 9.8% der Teilnehmer angaben, folgen anschließend nur noch Gruppen, welche maximal 1 bis 2 mal vorkamen.

3.3.4. Gemeinsames Auftreten von Gruppen

Das gemeinsame Auftreten von Kontaktgruppen wurde in Abbildung 3.3 durch einen Graphen visualisiert. Die Knoten repräsentieren die jeweiligen Kontaktgruppen, während Kantengewichte angeben, wie oft dazu inzidente Kontaktgruppen zusammen von einem Teilnehmer genannt wurden. Da eine Visualisierung aller von Teilnehmern genannten Kontaktgruppen zur Unübersichtlichkeit führt, wurden in Abbildung 3.3 nur die ersten acht in Tabelle 3.2 genannten Kontaktgruppen berücksichtigt.

Zu erkennen ist, dass die Verbindung zwischen den Gruppenbezeichnern *Familie* und *Freunde* mit einer Gewichtung von 52 die stärkste ist. 49 mal wurde die Kontaktgruppe *Familie* zusammen mit der Kontaktgruppe *Arbeit* genannt. Die drittstärkste Verbindung mit einer Gewichtung von 44 besteht zwischen den Kontaktgruppen *Arbeit* und *Freunde*n. Das geringste Gewicht einer Kante liegt bei 4 gemeinsam genannten Gruppen. Diese verbinden

Rang	Gruppenbezeichner	Vorkommen
1	Familie	65 (79.3%)
2	Freunde	58 (70.7%)
3	Arbeit	46 (56.1%)
4	Uni	27 (32.9%)
5	Andere	18 (22.0%)
6	Bekannte	18 (22.0%)
7	Hobby	9 (11.2%)
8	Enge Freunde	8 (9.8%)
...	...	...
32	Privat	1 (1.2%)

Tabelle 3.2.: Bevorzugte Kontaktgruppen für die Gruppierung von Kontakten (Auswertung zu Frage 4)

jeweils die Kontaktgruppen *Uni* mit *Hobby*, *Uni* und *Familie*, *Freunde* mit *Enge Freunde* sowie *Hobby* und *Andere*.

Kontaktgruppen, welche nicht durch eine Kante miteinander verbunden sind, kamen in keiner Antwort gemeinsam vor. Dies wäre beispielsweise der Fall bei den Kontaktgruppen *Bekannte* und *Hobby* oder *Bekannte* und *Andere*.

3.3.5. Anzahl benötigter Kontaktgruppen für die Gruppierung

Tabelle 3.3 stellt die Anzahl benötigter Kontaktgruppen in Bezug auf die Größe der Kontaktliste der Teilnehmer dar. Durchschnittlich werden bei Teilnehmern, die weniger als 50 Kontakte auf ihrem Smartphone besitzen, 3.65 Gruppen für die Gruppierung der Kontakte verwendet. Von durchschnittlich 4.07 Kontaktgruppen machen Teilnehmer, die zwischen 50 und 100 Smartphone-Kontakte besitzen, Gebrauch. Teilnehmer, welche zwischen 100 und 200 Kontakte auf ihrem Smartphone speichern, benutzen durchschnittlich 4.36 Kontaktgruppen für die Gruppierung.

Alle Teilnehmer, die angaben mehr als 200 Kontakte zu besitzen, gaben neben standardmäßigen Gruppenbezeichner auch generische Vertreter, wie beispielsweise <*projects*> oder <*country*>, an. Dies bedeutet, dass diese Teilnehmer für jedes ihrer Projekte bzw. für jedes von ihnen besuchte Land eine Kontaktgruppe angelegt haben.

Zählt man generische Kontaktgruppen als 2 oder mehr, so stellt sich heraus, dass die Anzahl der zur Kategorisierung benötigten Kontaktgruppen mit der Anzahl an Kontakten steigt.

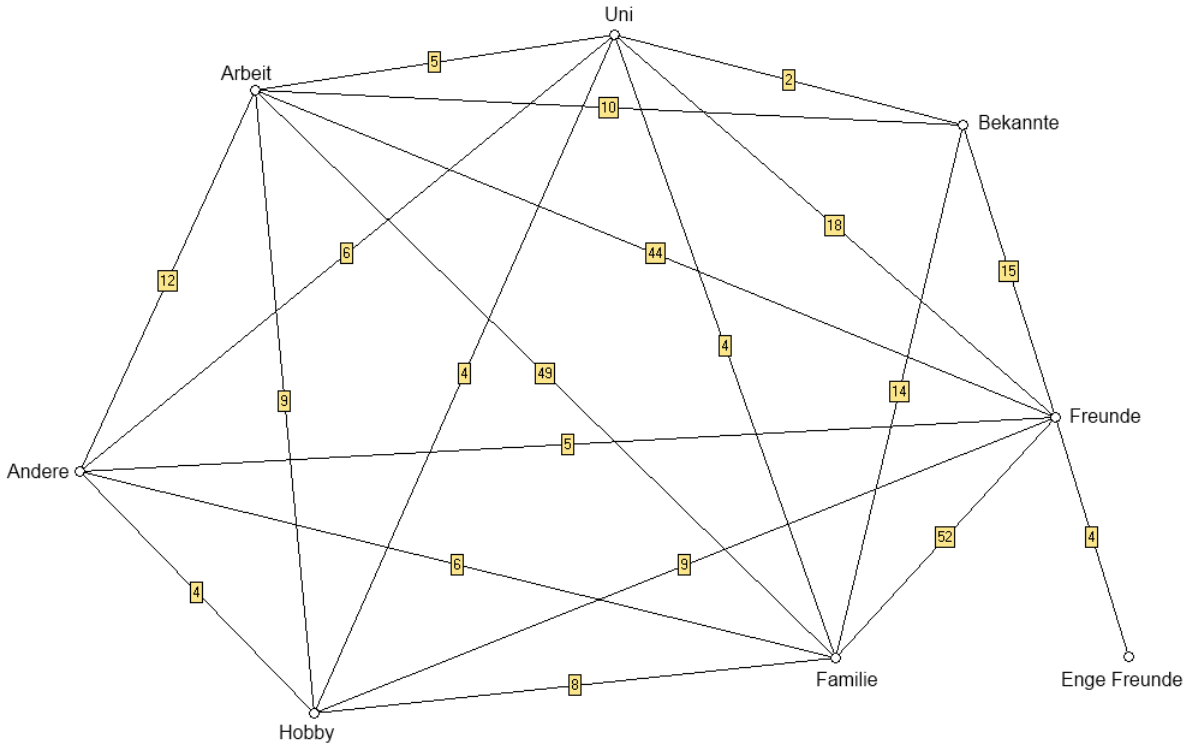


Abbildung 3.3.: Visualisierung gemeinsam vorkommender Kontaktgruppen in Antworten der Teilnehmer. Ein Knoten gibt die jeweilige Kontaktgruppe an, während das Kantengewicht die Anzahl an gemeinsamen Vorkommen repräsentiert. Eine vergrößerte Version wurde in Appendix A.2 abgedruckt.

Anzahl Kontakte	**Ø Anzahl an benötigten Kontaktgruppen**
weniger als 50	3.65
zwischen 50 und 100	4.07
zwischen 100 und 200	4.36
zwischen 200 und 400	4.0 (+ generische Gruppen)
mehr als 400	5.66 (+ generische Gruppen)

Tabelle 3.3.: Anzahl benötigter Kontaktgruppen in Abhängigkeit zur Größe der Kontaktliste. Generische Gruppen werden dabei als 2 oder mehr Gruppen gewertet.

3.4. Diskussion

Es wurde in diesem Kapitel ein Fragebogen erstellt, mit dem die Nutzungshäufigkeit von SMS und Anrufen, die Größe der Kontaktliste sowie die bevorzugten Kontaktgruppen zur Gruppierung der Kontaktliste von Probanden analysiert wurde. Eine Auswertung der Nutzungshäufigkeit von SMS und Anrufen zeigte, dass die Altersgruppe 18-25 Jahre bevorzugt SMS-Nachrichten zur Kommunikation verwenden. Dagegen gibt es bei der Altersgruppe der 26-50-Jährigen keine Vorzüge für SMS-Nachrichten oder Anrufen.

Aus einer Berechnung des gewichteten Durchschnitts der Größe der Kontaktliste in Bezug auf die Anzahl an Probanden folgt, dass alle Probanden durchschnittlich zwischen 100 und 200 Kontakten in ihren Kontaktlisten gespeichert haben. Daraus folgt die Annahme, dass, (wie in Tabelle 3.3 dargestellt) im Durchschnitt 4 bis 5 Kontaktgruppen für die Gruppierung der Kontaktliste benötigt wird.

Eine nähere Betrachtung der Kontaktgruppen *Hobby* und *Uni* im Graphen von Abbildung 3.3 zeigt, dass die genannten Kontaktgruppen keine inzidente Kante zur Kontaktgruppe *Bekannte* besitzen. Das heißt, dass unter den Befragten die Kontaktgruppe *Bekannte* nur dann vorkam, wenn die Gruppe *Hobby* nicht schon genannt worden ist und umgekehrt. Da Kontakte, mit denen die Teilnehmer lediglich wegen eines bestimmten Hobbys (viele gaben bspw. einen Sportverein an) in Kontakt bleiben, keine Freunde in dem Sinne sind, ist davon auszugehen, dass mit *Hobby* Kontakte gemeint sind, zu denen Teilnehmer einen bekanntschaftlichen Kontakt pflegen. Ähnliches Verhalten gilt auch für die Kontaktgruppe *Uni*. Aus diesem Grund werden für die weitere Arbeit alle Kontakte der Gruppe *Hobby* sowie *Uni* in die Gruppe *Bekanntschaft* aufgenommen.

Nach Zusammenfassung der Gruppen *Hobby* und *Uni* mit den *Bekannten* und der Ermittlung der Anzahl an benötigten Kontaktgruppen folgen daraus abschließend folgende fünf Kontaktgruppen, welche in der weiteren Arbeit als *vorgegebene Kontaktgruppen* verwendet und bezeichnet werden:

- Familie *(von 79.3 % aller Teilnehmer genannt)*
  Familienangehörige wie Vater, Mutter, Schwester, Bruder, Frau, Mann, etc.
- Freunde *(von 70.7 % aller Teilnehmer genannt)*
  Zu diesen Kontakten besteht ein regelmäßiger bis enger Kontakt.
- Bekannte *(von 65.5 % aller Teilnehmer genannt)*
  Zu diesen Kontakten besteht wenig Kontakt.
- Arbeitskollegen *(von 56.1 % aller Teilnehmer genannt)*
  Arbeitskollegen, Kunden, Chef, etc.
- Andere *(von 22.0 % aller Teilnehmer genannt)*
  Servicenummern wie bspw. Hausarzt, ADAC, Pizzaservice oder Taxi.

4. System zur Sammlung von Kontakt- und Aktivitätsdaten

Nachdem im letzten Kapitel die vorgegebenen Kontaktgruppen sowie weitere Informationen ermittelt wurden, wird in diesem Kapitel nun ein System entwickelt, mit dem Daten auf Smartphones im Rahmen einer Studie gesammelt werden. Dabei bestehen diese Daten, welche im Folgenden mit *Kontakt- und Aktivitätsdaten* bezeichnet werden, aus Kontakten mit ihrer SMS- und Anrufhistorie, sowie einer anschließenden manuellen Klassifizierung in die vorgegebenen Kontaktgruppen durch die Probanden. Die Studie, in der die entwickelte Anwendung benötigt wird, wird nach diesem Kapitel durchgeführt. Die dadurch gesammelten Daten werden anschließend ausgewertet und für die Entwicklung des Klassifikators eingesetzt. Somit bilden diese zu sammelnden Daten die Grundlage für die Entwicklung des Klassifikators.

In den folgenden Abschnitten wird die Architektur und die Funktionalitäten des Systems erläutert. Anschließend wird auf einige Aspekte der Implementierung eingegangen. Da das System für Probanden der Studie entwickelt wird, wird dabei der Anwender der Anwendung mit *Proband* bezeichnet.

4.1. Architektur

Das zu entwickelnde System besteht aus einer *Android-Anwendung* und einem *Webserver*. Die *Android-Anwendung* liest Kontakte inklusive ihrer SMS- und Anrufhistorie von Smartphones aus und lässt den Probanden anschließend seine Kontakte in die vorgegeben fünf Gruppen einordnen. Die so gesammelten Daten werden danach auf einen *Webserver* übertragen, sodass diese für die weitere Arbeit heruntergeladen und ausgewertet werden können.

Die Android-Anwendung wird nach der Entwicklung in Form einer APK-Datei [1] auf den *Webserver* hochgeladen. Die Distribution an die Probanden erfolgt durch einen QR-Code [2], welcher die URL zur APK-Datei beinhaltet. Die zu entwickelnde Android-Anwendung soll auf Smartphones mit Android 2.1 oder aktueller lauffähig sein. Auf dem Webserver sind eine PHP-Installation [3] sowie MySQL-Datenbanken [4] vorhanden.

[1] APK (**A**ndroid **P**ac**k**age) ist ein Dateiformat, mit der Android-Anwendungen installiert werden können

[2] Das QR steht dabei für Quick Response. Mithilfe einer Anwendung kann der darin enthaltene Text ausgelesen werden.

[3] *PHP Hypertext Preprocessor*: http://www.php.net/ *(zuletzt geprüft: 25.11.2012)*

[4] *Website der MySQL-Datenbank*: http://www.mysql.de/ *(zuletzt geprüft: 25.11.2012)*

4.2. Funktionalität

In diesem Abschnitt werden die Komponenten des Systems anhand ihrer Funktionalitäten beschrieben. Dazu wird zunächst die Android-Anwendung zum Auslesen der Kontakt- und Aktivitätsdaten und anschließend der Webserver beschrieben.

Android Anwendung zum Auslesen der Kontakt- und Aktivitätsdaten

Nach dem Starten der Anwendung hat der Proband die Möglichkeit sich für die Sprachen Deutsch oder Englisch zu entscheiden. Dies erlaubt auch die Rekrutierung englischsprachiger Probanden für diese Studie.

Anschließend erfolgt das Auslesen der Kontakte mit ihrer SMS- und Anrufhistorie. Dabei werden ausschließlich Kontakte berücksichtigt, mit denen der Proband bereits in Form von SMS-Nachrichten oder Anrufen kommuniziert hat. In diesem Schritt werden maximal 50 Kontakte ausgelesen, da sonst ein Proband, welcher mit überdurchschnittlich vielen Kontakten auf seinem Smartphone kommuniziert, mehr Kontakte (und somit auch Klassifizierungen) für die weitere Arbeit zur Verfügung stellt als ein Proband, auf den dies nicht zutrifft. Dies würde dazu führen, dass Probanden, die viele Kontakte zur Verfügung stellen, einen größeren Einfluss auf die weitere Arbeit ausüben als Probanden, auf deren Smartphone dagegen wenig Kontakte zu finden sind. Den Fall betreffend, bei dem mehr als 50 Kontakte mit SMS- und Anrufhistorie zum Auslesen zur Verfügung stehen, werden die auszulesenden Kontakte mit dem in Abschnitt 4.3.1 beschriebenen Algorithmus gewählt. Sind dagegen keine solcher Kontakte auf dem Smartphone zu finden, wird die Anwendung mit einem Warnhinweis beendet.

Bevor im nächsten Schritt die Klassifizierung der Kontakte durch den Probanden erfolgt, werden zunächst in Form eines Fragebogens demographische Informationen über den Probanden erhoben. Zu diesen Informationen gehören neben Alter und Geschlecht auch die Informationen ob der Proband neben SMS-Nachrichten andere Nachrichtendienste verwendet sowie die Verwendungshäufigkeit von SMS und Anrufen.

Nach dem Auslesen der Kontakte und dem Erheben demographischer Informationen folgt im nächsten Schritt die Klassifizierung der Kontakte durch den Probanden. Dies geschieht durch eine Bildschirmmaske, auf der nacheinander die ausgelesenen Kontakte anhand ihres eingespeicherten Namens, ihrer Telefonnummer sowie gegebenenfalls ihres Kontaktbildes angezeigt werden. (s. Abbildung 4.1) Zu jedem Kontakt werden die fünf vorgegebenen Kontaktgruppen angezeigt, in die der Proband den angezeigten Kontakt einordnet.

Hat der Proband mit der Klassifizierung seiner Kontakte abgeschlossen, versucht die Anwendung nun eine Verbindung zum *Webserver* aufzubauen. Ist dies möglich, so werden in diesem Schritt die gesammelten Daten übertragen. Ist auf dem Smartphone jedoch keine Internetverbindung verfügbar, so werden die gesammelten Daten auf dem Datenspeicher des Smartphones gesichert. Diese können später manuell vom Smartphone auf den Computer

Abbildung 4.1.: Screenshot der Android-Anwendung, die für die Sammlung von Kontakt- und Aktivitätsdaten zuständig ist.

übertragen werden. Die gesammelten Daten werden im JSON-Format gespeichert. Dessen Aufbau wird in Abschnitt 4.3.2 detaillierter erläutert.

Webserver

Auf dem Webserver werden sowohl die Android-Anwendung in Form einer APK-Datei sowie die gesammelten Daten gespeichert. Die Android-Anwendung wird dabei auf das Dateisystem des Webserver übertragen und anhand einer URL in Form eines QR-Codes an Probanden verteilt.

Die von der Anwendung gesammelten Daten werden dagegen in einer Datenbank auf dem Webserver gespeichert. Dazu bietet der Server ein PHP-Skript an, an die beim Aufruf durch POST-Parameter die Daten mitgeliefert werden können. Die so mitgelieferten Daten werden anschließend vom PHP-Skript in die MySQL-Datenbank geschrieben.

4.3. Implementierung

In diesem Abschnitt wird die Implementierung des Systems genauer erläutert. Dabei werden auf relevante Funktionalitäten des Systems sowie auf die Struktur der im JSON-Format gespeicherten Daten eingegangen. Der Quellcode des Systems wird als Anhang zu dieser Arbeit mitgeliefert.

Die Android-Anwendung wurde mithilfe des Android SDK in der Version 20.0.3 und Eclipse 4.2 Juno entwickelt. Als Testgeräte für die Entwicklung dienten ein Samsung Galaxy Nexus, ein Motorola Defy und der mit dem Android SDK mitgelieferte Emulator. Bevor die Studie begann, wurde die Anwendung im engeren Freundeskreis ausführlich getestet.

4.3.1. Funktionalitäten des Systems

Im Folgenden wird auf die Implementierung relevanter Funktionalitäten des Systems eingegangen. Zu den relevanten Funktionalitäten der Android-Anwendung gehören das Auslesen der Kontakte mit ihrer SMS- und Anrufhistorie, die im letzten Abschnitt angesprochene verhältnisgleiche Aussortierung sowie die Schnittstelle zwischen der Android-Anwendung und dem Webserver.

Auslesen von SMS Nachrichten und Anrufen

SMS Nachrichten sowie Anrufe werden in Android mithilfe zweier ContentProvider ausgelesen. Die URIs der ContentProvider lauten wie folgt:

- Anrufe: `content://call_log/calls` oder als vordefinierte Konstante in `CallLog.Calls.CONTENT_URI`
- SMS Nachrichten: `content://sms`
  Für den SMS ContentProvider gibt es aktuell weder eine vordefinierte Konstante noch eine Dokumentation in der API. Nach inoffiziellen Quellen [5] und [6] wurden diese entfernt, um Geräteherstellern die Möglichkeit zu geben die Standard Anwendung für SMS durch ihre eigene zu ersetzen, welche eingehende SMS Nachrichten sowie mögliche Zusätze an einem anderen Ort speichern. Um Probleme zu vermeiden, wurde die Anwendung vor Durchführung der Studie auf zahlreichen Smartphones im Freundeskreis getestet.

Um Daten aus einem ContentProvider zu lesen, wird zunächst ein Cursor auf die Daten über den *ContentResolver* [7] der Activity angefordert. Im folgenden Beispiel in Listing 4.1 geschieht dies in den Zeilen 1 bis 9. Dabei wird der Methode query() des Cursors die URI für die Anrufhistorie, sowie vier weitere Parameter übergeben. Da der Cursor in diesem Beispiel mit den Cursorn aus Datenbanksystemen [KE04] vergleichbar sind, bestehen die weiteren vier Parameter aus der Projektion in Form von Spalten des ContentProviders, aus der Selektion und ihre Argumente sowie der Sortierung der gelesenen Datensätze. Im nächsten Schritt wird in Zeile 11 mit der Methode moveToNext() des Cursors über alle

[5] *Be careful with content providers*: http://android-developers.blogspot.de/2010/05/be-careful-with-content-providers.html *(zuletzt geprüft: 25.11.2012)*
[6] *API for reading SMS messages removed from 1.0*: http://code.google.com/p/android/issues/detail?id=914 *(zuletzt geprüft: 25.11.2012)*
[7] *Android SDK (ContentResolver)*: http://developer.android.com/reference/android/content/ContentResolver.html *(zuletzt geprüft: 25.11.2012)*

ausgelesenen Datensätze iteriert. Anhand eines Index wird in den Zeilen 15-17 auf die Spalten der ausgelesenen Datensätze zugegriffen.

```
1  ContentResolver cr = getContentResolver(); // ContentResolver von Activity erhalten
2  Cursor c = cr.query(CallLog.Calls.CONTENT_URI, // URI
3                                  new String[] { CallLog.Calls.NUMBER, // Spalten
4                                         CallLog.Calls.DATE,CallLog.Calls.DURATION
5                                  },
6                                  null, // Selektion
7                                  null, // Selektionsargumente
8                                  null  // Sortierung
9                            );
10
11 while (c.moveToNext()) {
12        // ..
13        // .. Verarbeitung oder Speicherung erfolgt hier
14        // ..
15        String number = c.getString(0);
16        String duration = c.getString(2);
17        String timestamp = c.getString(1);
18 }
19 c.close();
```

Listing 4.1: Beispiel für das Auslesen von Daten aus einem ContentProvider.

Auslesen von Kontakten

Auch das Auslesen von sonstigen Kontaktinformationen wie Name oder Kontaktbild erfolgt über ContentProvider. Um Informationen zu einem bestimmten Kontakt oder einer bestimmten Telefonnummer zu erhalten, kann, wie in 5.2.1 beschrieben, eine Selektion angewendet werden. Dabei werden alle Kontakte ausgefiltert, welche nicht der übergebenen Kontakt-ID oder Telefonnummer entsprechen.

Für den Kontaktnamen, dem Kontaktfoto sowie der Kontakt-ID wurde die folgende Konstante als ContentProvider URI verwendet: `ContactsContract.Contacts.CONTENT_URI`. Die Spalten, welche in Form eines Arrays als Parameter für `projection` übergeben werden, lauten: `Contacts.DISPLAY_NAME`, `Contacts.PHOTO_ID` und `Contacts.CONTACT_ID`.

Verhältnisgleiche Aussortierung

Um einen zu großen Einfluss auf die Studie durch Probanden mit überdurchschnittlich vielen Kontakten zu vermeiden, sollen so viele Kontakte gestrichen werden, bis maximal 50 Kontakte übrig bleiben. Sollen mehrere Kontakte aussortiert werden, so muss darauf geachtet werden, dass diese gleichverteilt aus der Liste gestrichen werden. Das heißt, dass sowohl Kontakte, mit denen viel kommuniziert wird, als auch Kontakte, mit denen wenig kommuniziert wird, gleichermaßen aus der Liste gestrichen werden. Dadurch soll verhindert werden, dass im Falle der Verwendung der verhältnisgleichen Aussortierung mehr Kontakte,

die viel mit den Probanden kommunizieren, ausgelesen werden als welche, die weniger mit den Probanden in Kontakt treten.

Diese Aufgabe soll folgender Algorithmus lösen:

```
1  int maxSize = 50;
2  f = (contactlist.size() / maxSize) - 1;
3  error = 0;
4  i = 0;
5
6  while (i < contactlist.size()) {
7         if (error < 1) {
8                 error += f;
9                 i++;
10        } else {
11                error = error - 1;
12                contactlist.remove(i);
13        }
14 }
```

Listing 4.2: Verhältnisgleiche Aussortierung im Falle von mehr als 50 verfügbaren Kontakten zum Auslesen

Zu Beginn wird in Zeile 2 der Faktor f berechnet, der angibt, um welchen Faktor die aktuelle Liste zu viel an Kontakten in Bezug auf die erwartete Größe hat. Anschließend wird in Zeile 6 eine Iteration durch die Liste geführt. Mit jedem Kontakt, der in der Liste behalten wird, steigt der error-Wert in Zeile 8 inkrementell um den Faktor f. Wird der error-Wert bei einer Iteration größer als 1, so wird der aktuell betrachtete Kontakt aus der Liste gestrichen. (Zeile 12) Der error-Wert wird anschließend um 1 subtrahiert. Diese Iteration wird so lange durchgeführt, bis alle Kontakte der Liste betrachtet wurden. Dieser Algorithmus erlaubt dadurch eine gleichmäßige Entfernung von Kontakten, wobei der Abstand zwischen diesen Kontakten, bedingt durch den error-Wert, in nahezu allen Fällen gleich groß ist.

Übertragen der Daten von der Android-Anwendung auf den Webserver

In der Android-Anwendung wurde das Hochladen der Daten durch die Java Klassen *HttpClient* sowie *HttpPost* realisiert. Beide Klassen sind im Paket `org.apache.http.client` zu finden, das Java standardmässig mitliefert.

Die Klasse *HttpClient* stellt eine einfache HTTP Verbindung zu einem Server her, mit der Daten gesendet und empfangen werden können. Mithilfe der Klasse *HttpPost* können zu einer HTTP Verbindung zusätzlich Wertetupel (in diesem Fall POST-Parameter) übertragen werden.

Anhand dieser Klassen wird eine Verbindung zum Webserver über ein PHP-Skript hergestellt. Diesem PHP-Skript werden dabei anhand der POST-Parameter die gesammelten Daten übertragen. Anschließend werden die übertragenen Daten durch das PHP-Skript in die MySQL-Datenbank geschrieben.

4.3.2. Format zur Speicherung der gesammelten Daten

Ausgelesene Daten werden im JSON-Format gespeichert. Zu jedem Probanden wird ein JSON-Objekt gespeichert, das die Probanden-ID, die Antworten zum Fragebogen sowie eine Liste von Kontakten beinhaltet. Zu einem Kontakt wird eine Kontakt-ID, die zur Klassifizierung gewählten Kontaktgruppen sowie zwei Listen gespeichert. In den zwei Listen wiederum werden Informationen zu getätigten Anrufen bzw. gesendeten/empfangenen SMS gespeichert. Zu diesen Informationen gehören die Länge des Anrufs bzw. der Textnachricht sowie der Zeitpunkt in Form eines Timestamps.

4.4. Zusammenfassung

Es wurde in diesem Kapitel ein System entwickelt, mit der in der weiteren Arbeit Daten von Probanden im Rahmen einer Studie gesammelt werden. Zu diesen Daten gehören die Smartphone-Kontakte mit ihrer SMS- sowie Anrufshistorie sowie die Antworten zu einigen demographischen Fragen. Dazu wurde eine Android-Anwendung entwickelt, welches für das Auslesen und Übertragen der genannten Daten zuständig ist. Auf dem Webserver werden anschließend die ausgelesenen Daten gespeichert. Zudem wird auf dem Webserver die Android-Anwendung in Form einer APK-Datei zum Download verfügbar sein. Für die Distribution der Anwendung wurde ein QR-Code generiert, welches die URL zur APK-Datei beinhaltet.

5. Sammlung und Auswertung von Kontakt- und Aktivitätsdaten

In den letzten Kapiteln wurden zunächst die bevorzugten Kontaktgruppen ermittelt und anschließend ein System zur Sammlung von Kontakt- und Aktivitätsdaten entwickelt. Dieses System wird in diesem Kapitel nun eingesetzt, um die für die weitere Arbeit notwendigen Daten im Rahmen einer Studie zu sammeln. Nach der Sammlung werden die Daten ausgewertet, um daraus eine Grundlage für die Entwicklung des Klassifikators im nächsten Kapitel zu bilden.

5.1. Methodik

Nach Fertigstellung des Systems zur Sammlung von Kontakt- und Aktivitätsdaten wurde die Android-Anwendung des Systems auf einem Webspace hochgeladen. Mithilfe eines QR-Codes konnten Probanden die Anwendung herunterladen und auf ihrem Smartphone installieren. Die Verwendung des Google Play Stores [1] wurde bewusst vermieden, da auf diesem Wege fremde Nutzer die Anwendung hätten installieren und somit möglicherweise das Ergebnis verfälschen können.

5.1.1. Ablauf

Die Studie wurde persönlich mit Studenten und Mitarbeitern der Universität Stuttgart durchgeführt. Die Rekrutierung der Probanden erfolgte mündlich. Als Bedingung für eine Teilnahme wurde ein Smartphone angegeben, auf dem Android 2.1 oder aktueller läuft. Zudem sollen Probanden ihr Smartphone bereits länger als einen Monat besitzen und schon mit ihren Kontakten über dieses kommuniziert haben. Selbstverständlich durften die SMS-Historie und Anrufhistorie nicht gelöscht werden. Die Studie wurde fünf Wochen lang vom 01.07.2012 bis zum 08.08.2012 durchgeführt.

Vor Beginn der Studie bekamen die Probanden eine Einführung in die Studie sowie eine Erklärung der fünf vorgegebenen Kontaktgruppen. Nach Beginn der Studie durften Probanden keine Fragen mehr zu den Gruppen stellen, da dies sonst Einflüsse auf die Entscheidungen haben könnte.

[1] *Google Play Store*: http://play.google.com *(zuletzt geprüft: 25.11.2012)*

5.1.2. Teilnehmer

An dieser Studie nahmen 20 Probanden teil, die jeweils zwischen 18 und 27 Jahre alt waren. Davon waren 12 Teilnehmer weiblich und 8 männlich. 14 Teilnehmer gaben an, dass sie neben SMS-Nachrichten auch andere Nachrichtendienste wie beispielsweise WhatsApp verwenden. Der Großteil aller Probanden bestand aus Studenten oder Mitarbeiter an der Universität Stuttgart.

Da nur Kontakte ausgelesen werden, für die eine SMS-Historie oder eine Anrufhistorie existiert, und bei jedem Proband maximal 50 Kontakte ausgelesen werden sollten, kamen für diese Studie insgesamt 439 Kontakte zusammen. Im Durchschnitt wurden von jedem Probanden 22 Kontakte ausgelesen (min = 6, max = 50, stdev. = 10,49). Der älteste Eintrag in den ausgelesenen SMS- und Anrufhistorien datierte im Februar 2011.

5.2. Ergebnisse

Nachdem im letzten Abschnitt die Kontakt- und Aktivitätsdaten gesammelt wurden, werden diese in diesem Abschnitt nun ausgewertet. Hierfür wurde ein Programm entwickelt, welches die im JSON-Format gesammelten Daten einliest und daraus die im Folgenden vorgestellten Auswertungen erstellt. Dieses Programm wird ab jetzt als *Hilfsprogramm* bezeichnet. Zudem wird im Folgenden eine versendete oder empfangene SMS, beziehungsweise ein eingehender oder ausgehender Anruf als *Aktivität* bezeichnet.

Verteilung der Kontaktgruppen Abbildung 5.1 zeigt die Verteilung der vorgegebenen Kontaktgruppen für die Gruppierung der Kontakte. Deutlich zu erkennen ist, dass bei den insgesamt 439 Kontakten die Kontaktgruppen *Freunde* und *Bekannte* am häufigsten für die Gruppierung gewählt wurden. So wurden 153 Kontakte als *Bekannte* klassifiziert, während 139 in die Gruppe der *Freunde* eingeordnet wurden. 77 Kontakte wurden als *Familienmitglieder* gekennzeichnet. Mit 43 Kontakten in der Gruppe *Andere* und 27 Kontakten in der Gruppe der *Arbeitskollegen* sind dies die am wenigsten verwendeten Kontaktgruppen.

Aktivitäten einzelner Probanden Abbildung 5.2 zeigt die Aktivität aller 20 Probanden auf einen Blick. Die Abszissenachse (x-Achse) gibt die jeweiligen Probanden anhand von Nummern (insgesamt 20) an. Auf der Ordinatenachse (y-Achse) ist die Aktivität in 200er Schritten gemessen. Dabei ist zu erkennen, dass bei nahezu allen Probanden die SMS-Aktivität deutlich über den Telefonaten liegt. Die SMS-Aktivität von Proband 6 stellt im Diagramm einen Ausreißer dar, welcher mit 30891 SMS-Nachrichten nicht in das Diagramm mit aufgenommen werden konnte. Alle 20 Probanden tätigten insgesamt 3895 Anrufe und 37189 SMS-Nachrichten.

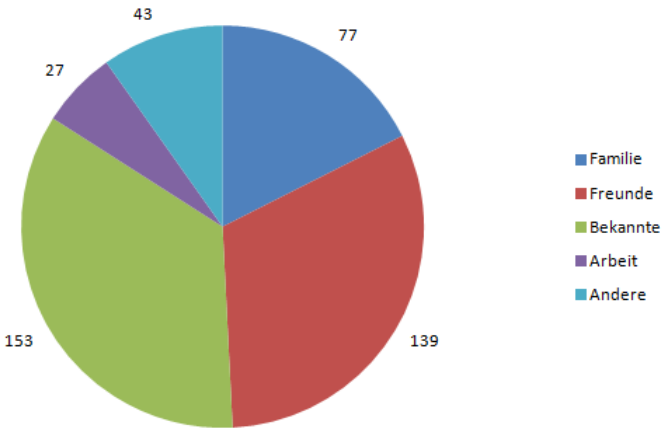


Abbildung 5.1.: Verteilung der vorgegebenen Kontaktgruppen für die Gruppierung von Kontakten durch Probanden.

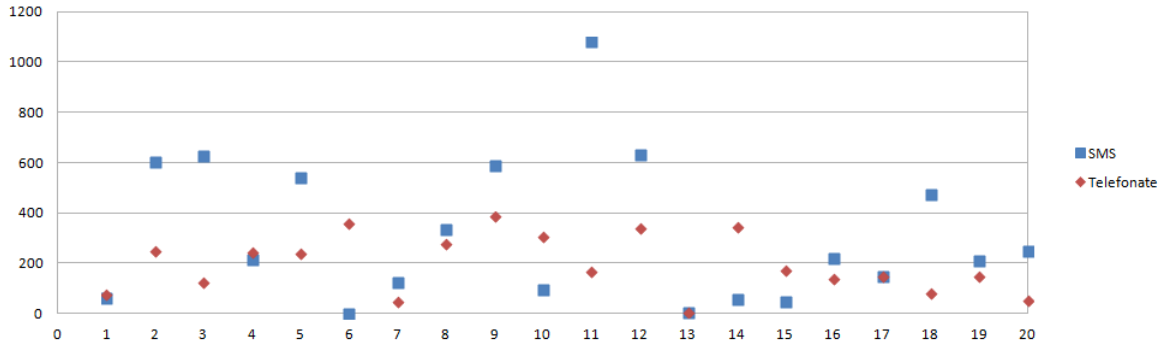


Abbildung 5.2.: Übersicht über die Anzahl an SMS-Nachrichten und Anrufen aller Probanden. Dabei wird auf der Abszissenachse die Probanden anhand von Nummern dargestellt, während auf der Ordinatenachse die Anzahl an SMS-Nachrichten und Anrufen in 200er Schritten gemessen wird.

SMS und Anrufe Tabellen 5.1 und 5.2 zeigen SMS- und Anrufaktivitäten jeweiliger Gruppen anhand der durchschnittlichen Anzahl an Aktivitäten sowie die Minima, Maxima und Standardabweichungen pro Gruppe.

Die Kontaktgruppe *Familie* versendet im Durchschnitt 3.3 SMS-Nachrichten und tätigt durchschnittlich 13.9 Anrufe. Aus der Kontaktgruppe *Familie* wurden dabei maximal 161 Anrufe pro Kontakt getätigt und 49 SMS-Nachrichten pro Kontakt ausgetauscht. Durchschnittlich tauschen Kontakte der Gruppe *Freunde* 236.9 (stdev = 778.02) SMS-Nachrichten aus und führen durchschnittlich 16.31 Telefonate. Auffällig ist dabei ein Kontakt der Gruppe, welcher zum Zeitpunkt der Studie bereits 5000 SMS-Nachrichten ausgetauscht hat. Dagegen kommen

Gruppenname	Ø Anzahl SMS *(Std. Dev)*	Min. pro Kontakt	Max. pro Kontakt
Familie	3.3 *(8.66)*	0	49
Freunde	236.9 *(778.02)*	0	5000
Bekannte	18.3 *(77.06)*	0	775
Arbeit	4.1 *(8.62)*	0	39
Andere	0.1 *(0.21)*	0	1

Tabelle 5.1.: Anzahl an SMS-Aktivitäten jeweiliger Kontaktgruppen. Dabei bezieht sich die durchschnittliche Anzahl an SMS (sowie die Standardabweichung) auf alle Kontakte der jeweiligen Gruppe, während sich das Minimum und Maximum auf einen einzelnen Kontakt einer Gruppe bezieht.

Gruppenname	Ø Anzahl Anrufe *(Std. Dev)*	Min. (pro Kontakt)	Max. (pro Kontakt)
Familie	13.9 *(23.86)*	1	161
Freunde	16.31 *(25.98)*	0	159
Bekannte	1.84 *(3.53)*	0	29
Arbeit	3.56 *(6.60)*	0	33
Andere	1.95 *(2.36)*	0	12

Tabelle 5.2.: Anzahl an Anruf-Aktivitäten jeweiliger Kontaktgruppen. Dabei bezieht sich die durchschnittliche Anzahl an Telefonaten (sowie die Standardabweichung) auf alle Kontakte der jeweiligen Gruppe, während sich das Minimum und Maximum auf einen einzelnen Kontakt einer Gruppe bezieht.

Kontakte der Gruppe *Bekannte* auf durchschnittlich 18.3 SMS-Nachrichten und 1.84 Anrufe. Ein Kontakt der Gruppe *Bekannte* tauschte dabei 775 SMS-Nachrichten zum Zeitpunkt der Studie aus. Bei 29 liegt die maximale Anzahl an Telefonaten, die ein Kontakt der Gruppe *Bekannte* versendet und empfangen hat. Die Kontaktgruppe *Arbeit* tauscht im Durchschnitt 4.1 SMS-Nachrichten aus, wobei der aktivste Kontakt 39 SMS-Nachrichten austauschte. Bis zu 3.56 Anrufen durchschnittlich werden im Vergleich zu den SMS-Nachrichten in der Kontaktgruppe *Arbeit* getätigt. Lediglich 0.1 SMS-Nachrichten durchschnittlich versendeten und empfingen die Kontakte der Gruppe *Andere* mit den Probanden aus. Dabei ist eine SMS pro Kontakt bereits das Maximum. Dagegen werden 1.95 Telefonaten in der genannten Kontaktgruppe getätigt, wobei das Maximum bei 12 Anrufen für einen Kontakt liegt.

Verhältnis von SMS zu Anrufen Abbildung 5.3 zeigt das Verhältnis zwischen den SMS-Aktivitäten und den Anruf-Aktivitäten für die vorgegebenen Gruppen. Bei den Gruppen *Familie* und *Andere* ist zu erkennen, dass SMS-Nachrichten mit jeweils 12.5% und 3.5% kaum Verwendung im Vergleich zu den Anrufen finden. Bei *Freunden* besteht der Kontakt zu 66.5% aus SMS-Nachrichten und 33.5% aus Anrufen. Ähnlich sieht dies bei den *Bekannten* mit 62.3% SMS und 37.7% Anrufen aus. Die Kontaktgruppe *Arbeit* schreibt zu 33% SMS und telefoniert zu 67%.

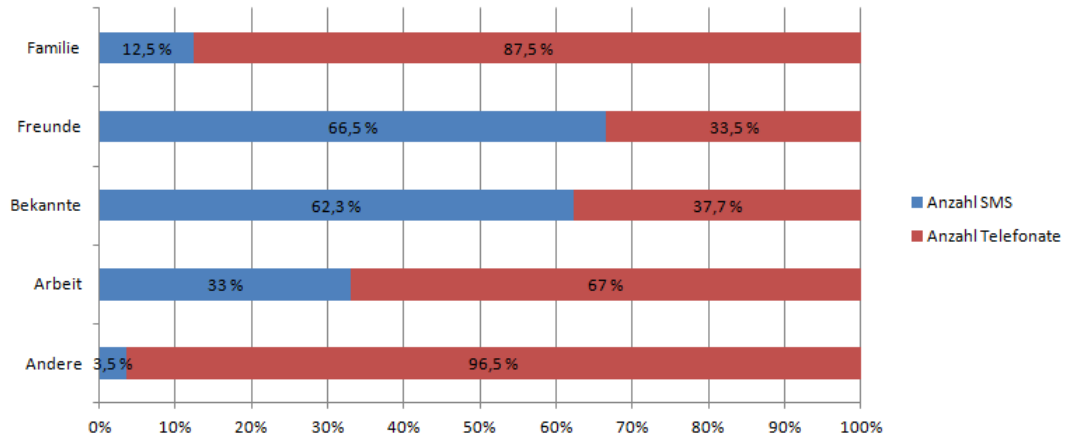


Abbildung 5.3.: Verhältnis zwischen SMS-Aktivitäten und Anruf-Aktivitäten der jeweiligen Gruppen. Die Angaben werden dabei prozentual dargestellt.

Es ist hier anzumerken, dass die Verhältnisse nicht für die Menge an ausgetauschten Informationen gelten. Mit einem Anruf können in den meisten Fällen mehr Informationen ausgetauscht werden als über eine SMS. Andersherum kann eine längere SMS mehr Informationen liefern als ein kurzer Anruf.

Länge von SMS-Nachrichten und Dauer von Anrufen Die Angaben in den Tabellen 5.3 und 5.4 repräsentieren jeweils die durchschnittliche SMS-Nachrichtenlänge in Anzahl an Zeichen beziehungsweise durchschnittliche Anrufdauer eines Kontakts in Sekunden. Die Spalten Minimum und Maximum beziehen sich somit auf die durchschnittlichen Längen von SMS oder Anrufen eines einzelnen Kontakt der jeweiligen Gruppe, während sich die Spalten *Ø aller Kontakte* und *Standard Deviation* aus allen Kontakten der Gruppe zusammensetzt.

Im Vergleich zur Anrufdauer gibt es bei der SMS-Nachrichtenlänge keine großen Unterschiede zwischen den Kontaktgruppen. Jeweils 16 bis 17 Zeichen beträgt die maximale Differenz zwischen der niedrigsten Durchschnittslänge aller Kontakte, welche die *Freunde* mit 73.41 Zeichen erreicht hat, und der größten Durchschnittslänge in der Gruppe *Andere* mit 90 Zeichen. Die Kontaktgruppe *Familie* und *Arbeit* schreiben mit durchschnittlich 80 Zeichen jeweils ähnlich lange SMS-Nachrichten. Die Kontaktgruppe der *Bekannten* schreibt durchschnittlich 89.06 Zeichen pro SMS, wobei es darunter einen Kontakt gibt, mit dem durchschnittlich 523 Zeichen pro SMS ausgetauscht wurden. Bei den *Freunden* kommt die Person, mit der die durchschnittlich längsten Nachrichten geschrieben werden, auf 257.55 Zeichen pro SMS. Im Vergleich dazu erreicht die Gruppe *Familie* ein Maximum von 260 Zeichen. Der pro Kontakt durchschnittlich höchste Wert in der Kontaktgruppe *Arbeit* liegt bei 158.5 Zeichen, in der Kontaktgruppe *Andere* beträgt dieser Wert 138.

Geht es um die Gesprächsdauer, so führen die Gruppen *Familie* und *Freunde* die längsten Telefonate. Durchschnittlich kommt die Gruppe *Familie* auf 87.45 Sekunden, *Freunde* kommen gar auf 89.80 Sekunden. Dagegen kommen die Kontaktgruppen *Bekannte*, *Arbeit* und *Andere*

Gruppe	Ø alle Kontakte *(Std. Dev)*	Minimum	Maximum
Familie	80.57 *(56.69)*	5.0	260.0
Freunde	73.41 *(39.82)*	17.17	257.55
Bekannte	89.06 *(70.68)*	9.0	523.0
Arbeit	80.95 *(42.98)*	2.0	158.5
Andere	90.0 *(67.88)*	42.0	138.0

Tabelle 5.3.: Durchschnittliche Länge von SMS-Nachrichten in den jeweiligen Gruppen. Dabei bezieht sich der Durchschnitt auf alle Kontakte jeweiliger Gruppen, während sich die Spalten Minimum und Maximum auf die durchschnittliche SMS-Länge eines einzelnen Kontakts in der jeweiligen beziehen. Die Einheit der Werte ist in Anzahl an Zeichen angegeben.

Gruppe	Ø alle Kontakte *(Std. Dev)*	Minimum	Maximum
Familie	87.45 *(134.22)*	0.0	736.57
Freunde	89.80 *(150.22)*	0.0	1025.79
Bekannte	42.05 *(147.34)*	0.0	1351.0
Arbeit	47.78 *(86.90)*	0.0	376.0
Andere	35.88 *(95.30)*	0.0	524.0

Tabelle 5.4.: Durchschnittliche Dauer aller Anrufe eines Kontakts. Dabei bezieht sich der Durchschnitt auf alle Kontakte jeweiliger Gruppen, während sich die Spalten Minimum und Maximum auf die durchschnittliche Dauer des Anrufs eines einzelnen Kontakts in der jeweiligen beziehen. Die Einheit der Werte wird in Sekunden angegeben.

lediglich auf Durchschnittswerte von 42.05 Sekunden, 47.78 Sekunden sowie 35.88 Sekunden. Bei den Anrufen ist somit deutlicher als bei den SMS-Nachrichten zu erkennen, welche Gruppen durchschnittlich länger telefonieren. Die längsten Telefonate mit einem Durchschnitt von 22 Minuten und 51 Sekunden wurden mit einem Kontakt aus der Kontaktgruppe *Bekannte* geführt, während der zweitlängste Durchschnitt auf einen Kontakt der Gruppe *Freunde* zurückzuführen ist.

Uhrzeiten Abbildungen 5.4 und 5.5 zeigen die Zeiten, an denen Kontakte der jeweiligen Gruppen SMS-Nachrichten ausgetauscht oder Anrufe getätigt haben. Dabei sind auf der Abszissenachse die Stunden abgebildet, an denen die jeweiligen Aktivitäten stattfanden. Die linke Ordinatenachse, die sich auf die Säulen des Diagramms bezieht, gibt das prozentuale Vorkommen der Gruppen an. Somit repräsentieren die Säulen prozentual das Vorkommen der jeweiligen Gruppen an einer bestimmten Stunde. Auf der rechten Ordinatenachse, die sich auf das Liniendiagramm bezieht, wird die normalisierte Aktivitätsanzahl gemessen. Diese Einheit wird im folgenden Absatz durch die Erstellung der Diagramme genauer erläutert.

Für die Erstellung der Diagramme wurde zunächst die Aktivität jedes einzelnen Kontakts für alle Stunden des Tages ermittelt. Damit jeder Kontakt unabhängig von seiner Gesamtaktivität den gleichen Einfluss auf das Ergebnis hat, wurden anschließend alle Aktivitäten durch einen Faktor f geteilt. Der Faktor f berechnet sich durch die Gesamtanzahl der Aktivitäten eines Kontakts geteilt durch 10. Somit ist sichergestellt, dass die Summe aller Aktivitäten eines Kontakts nach Division stets gleich 10 ist. Dadurch nimmt jeder Kontakt mit 10 Aktivitäten am Ergebnis teil, wobei das Verhältnis der Aktivitäten an den Stunden untereinander erhalten bleibt. Die Summe dieser Aktivitäten wird im folgenden *normalisierte Aktivitätsanzahl* genannt.

An der normalisierten Aktivitätsanzahl der SMS-Aktivitäten in Abbildung 5.4 ist zu erkennen, dass um 10 Uhr die meisten SMS-Nachrichten versendet und empfangen wurden. Dagegen wurden um 3 Uhr und 5 Uhr morgens die geringste SMS-Aktivität erkannt. Der Großteil an SMS-Aktivitäten, der mit einer normalisierten Aktivitätsanzahl von über 100 gekennzeichnet ist, fand zwischen 9 Uhr morgens und 23 Uhr statt, während von 0 Uhr bis 8 Uhr morgens nahezu kaum SMS-Nachrichten ausgetauscht werden. Lokale Hochpunkte der normalisierten Aktivitätsanzahl liegen hier bei 10 Uhr morgens, 14 Uhr sowie 21 Uhr. Bei den Anruf-Aktivitäten sind dagegen die lokalen Hochpunkte bei 11 Uhr morgens, 17 Uhr sowie 20 Uhr. Die meisten Anrufe wurden dabei um 17 Uhr getätigt. Der Großteil an Anruf-Aktivitäten wurde von 11 Uhr bis 23 Uhr durchgeführt.

Die Farbverteilung in Diagramm 5.4, die für die jeweiligen Kontaktgruppen steht, zeigt für die Zeiten von 7 Uhr bis 0 Uhr nahezu gleichgroße Anteile der Kontaktgruppen. Zwischen 3 und 5 Uhr sind ausschließlich Aktivitäten von Kontakten der Gruppe *Freunde* und *Bekannte* zu erkennen. Um 6 Uhr morgens wurde der größte Anteil an normalisierter Aktivitätsanzahl für die Kontaktgruppe *Familie* festgestellt. Die Kontaktgruppe *Andere* ist ausschließlich an den Zeiten 8 Uhr morgens und 17 Uhr zu erkennen. Anders verhält sich dies in Diagramm 5.5, welches die Anruf-Aktivitäten beschreibt. Die Kontaktgruppe *Andere* ist hier durchgehend zwischen 6 Uhr morgens und 0 Uhr zu erkennen. Dabei wurde um 6 Uhr morgens die meisten Anrufe der Gruppe durchgeführt. An allen Stunden des Tages außer 4 Uhr morgens, wurden Telefonate von der Kontaktgruppe *Familie* durchgeführt. Dabei wurden um 3 Uhr morgens die meisten Anrufe registriert. Im Gegensatz zu den SMS-Aktivitäten sind hier die Anteile der Kontaktgruppen *Freunde* und *Bekannte* zwischen 7 und 23 Uhr geringer.

5.3. Diskussion

In diesem Abschnitt werden die durch die Studie gewonnenen Daten diskutiert sowie Abhängigkeiten, die zu einer möglichen Unterscheidung der Gruppen beitragen, aufgeführt. Die hier gezeigten Abhängigkeiten sollen im nächsten Kapitel dazu genutzt werden, den Klassifikator zu entwickeln.

Verhältnis von SMS zu Anrufen In Abbildung 5.3 ist zu erkennen, dass die Kommunikation in den Gruppen *Freunde* und *Bekannte* größtenteils durch SMS-Nachrichten geführt wird.

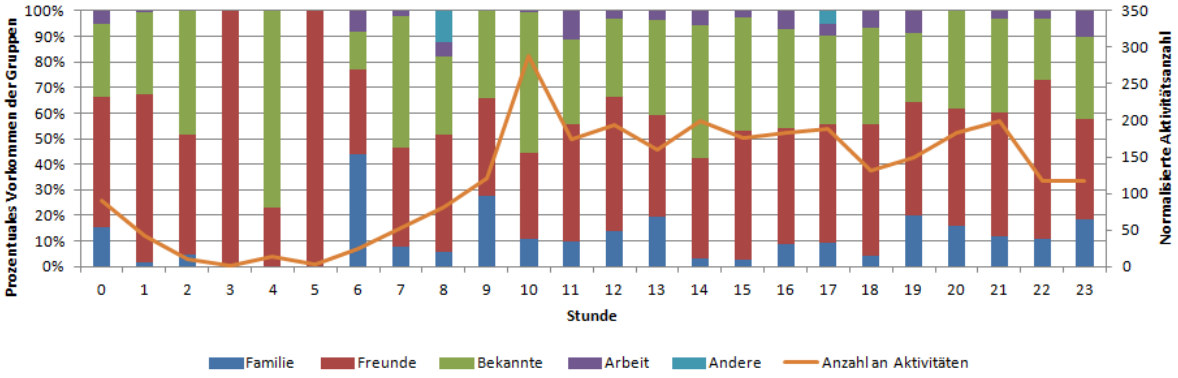


Abbildung 5.4.: Verteilung von SMS-Aktivitäten über den Tag. Auf der Abszissenachse werden die Stunden des Tages dargestellt, während auf der Ordinatenachse die normalisierte Aktivitätsanzahl in 50er Schritten gemessen wird. Bei der normalisierten Aktivitätsanzahl nimmt jeder Kontakt mit 10 Aktivitäten teil.

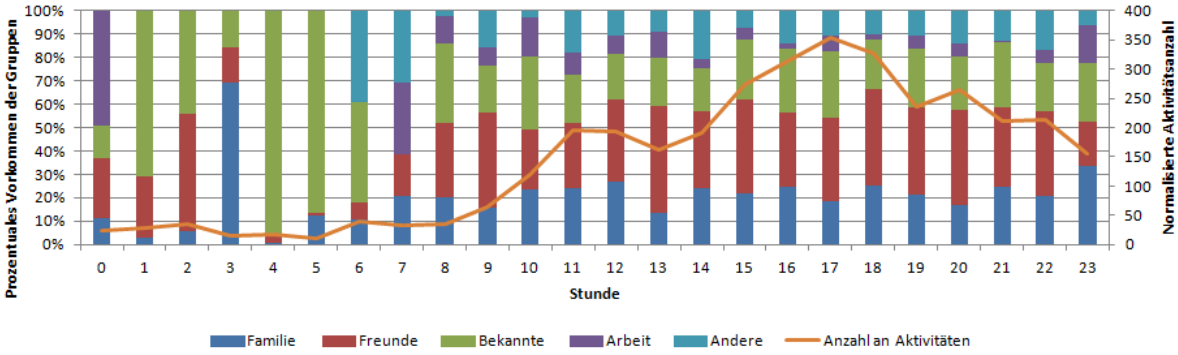


Abbildung 5.5.: Verteilung von Anruf-Aktivitäten über den Tag. Auf der Abszissenachse werden die Stunden des Tages dargestellt, während auf der Ordinatenachse die normalisierte Aktivitätsanzahl in 50er Schritten gemessen wird. Bei der normalisierten Aktivitätsanzahl nimmt jeder Kontakt mit 10 Aktivitäten teil.

In beiden Gruppen wird über 60% der Kommunikation mithilfe von SMS abgewickelt. Dagegen wird in den Gruppen *Familie*, *Arbeit* sowie *Andere* das Telefonieren bevorzugt. In der Kontaktgruppe *Andere* bestehen 96,5% der Kommunikation aus Anrufen, in der Gruppe *Familie* sind es 87,5%. Die Kontaktgruppe *Arbeit* zieht mit 67% das Telefonieren den SMS vor. Dieser Statistik zufolge ist eine Unterscheidung der Gruppen *Freunde* und *Bekannte* von den restlichen Gruppen anhand des Anruf-SMS-Verhältnisses durchaus denkbar.

Länge von SMS-Nachrichten und Dauer von Anrufen Durchschnittlich schreiben die Kontaktgruppen *Andere* und *Bekannte* die längsten SMS-Nachrichten. Die Längen der SMS-Nachrichten der Gruppen *Familie* und *Arbeit* befinden sich im mittleren Bereich, während

die Gruppe *Freunde* die kürzesten SMS-Nachrichten schreibt. Ein Vergleich der SMS-Längen mit der Anzahl ausgetauschter Nachrichten in Tabelle 5.1 zeigt, dass die SMS-Länge kleiner ist, wenn durchschnittlich mehr SMS Nachrichten versendet werden. Dies trifft auf alle Gruppen bis auf die der *Bekannten* zu. Da jedoch der Unterschied zwischen dem kürzesten und längsten Durchschnitt lediglich 16 Zeichen beträgt, ist dadurch eine Unterscheidung der Kontaktgruppen kaum möglich.

Die Gruppen *Familie* und *Freunde* führen durchschnittlich die längsten Telefonate mit annähernd 90 Sekunden. Die restlichen drei Gruppen kommen durchschnittlich auf 35-48 Sekunden. Hieran ist also deutlich zu erkennen, dass die Gruppen *Familie* und *Freunde* länger telefonieren als die Gruppen *Bekannte*, *Arbeit* und *Andere*. Denkbar wäre dadurch eine Abgrenzung der Gruppen *Familie* und *Freunde* von den restlichen Gruppen.

SMS- und Anrufranking Für jeden Probanden wurde eine Liste angelegt, in der die Kontakte absteigend nach SMS- und Anrufaktivität sortiert wurden. Anschließend wurde diese Liste in 8 gleichgroße Teillisten zerlegt, wobei die Teilliste mit dem ersten Achtel der Kontakte Rang 1 bekommt, das zweite Achtel Rang 2. Auch die restlichen 6 Teillisten werden mit diesem Verfahren den jeweiligen Rängen zugeordnet.

Abbildung 5.6 visualisiert das Auftreten der Gruppen in den 8 Rängen. In beiden Darstellungen ist festzustellen, dass Rang 1 und Rang 2 hauptsächlich von Kontakten der Gruppe *Freunde* besetzt werden, wohingegen in den niedrigeren Rängen vorrangig Kontakte der Gruppe *Bekannte* und *Andere* aufzufinden sind. Kontakte der Gruppe *Arbeit* sind am häufigsten in den Rängen 3 bis 6 vertreten. Geht es um SMS-Nachrichten, so erreicht das Vorkommen der *Familienmitglieder* erst in Rang 6 seinen Höhepunkt. Umgekehrt verhält sich die Kontaktgruppe *Familie* bei den Telefonaten. Hier sind im Rang 1 die meisten Kontakte der Kontaktgruppe *Familie* zu finden, während das Vorkommen mit jedem weiteren Rang abnimmt. Ab Rang 7 sind keine telefonierenden *Familienmitglieder* mehr vorhanden.

Anhand den Ranggruppen für SMS-Nachrichten und Anrufe lassen sich so verschiedene Gruppen voneinander unterscheiden. Als Beispiel ist hier die Gruppe *Freunde* und ihre SMS-Aktivität zu nennen. Während Rang 1 bis zu 85% aus *Freunden* besteht, ist ab dem 5. Rang nur noch ein Anteil von 7% vorhanden. Umgekehrt ist anzumerken, dass Kontakte der Gruppe *Andere* bei den SMS-Rankings nur in den Rängen 5 bis 7 häufig vorzufinden sind, wobei sie in allen anderen Rängen kaum einen Anteil besitzen. *Bekannte* können nahezu ausgeschlossen werden, wenn sich der Kontakt in der Rankinggruppe 1 für SMS sowie Anrufe befindet. Dagegen ist die Wahrscheinlichkeit für einen Kontakt der Gruppe *Bekannte* höher, wenn er sich in den Rängen 3-6 aufhält.

Uhrzeiten Aus den in den Abbildungen 5.4 und 5.5 dargestellten Diagrammen konnten keine Merkmale von Gruppen erkannt werden, anhand denen eine Unterscheidung von Gruppen möglich ist. Dies könnte möglicherweise an den gesammelten Kontakt- und Aktivitätsdaten liegen. Wird diese Auswertung für andere Kontakt- und Aktivitätsdaten durchgeführt, so wäre es sinnvoll die Zeiten in die Entwicklung der Klassifikatoren einzubeziehen, da diese eine mögliche Unterscheidung von Kontaktgruppen darstellen.

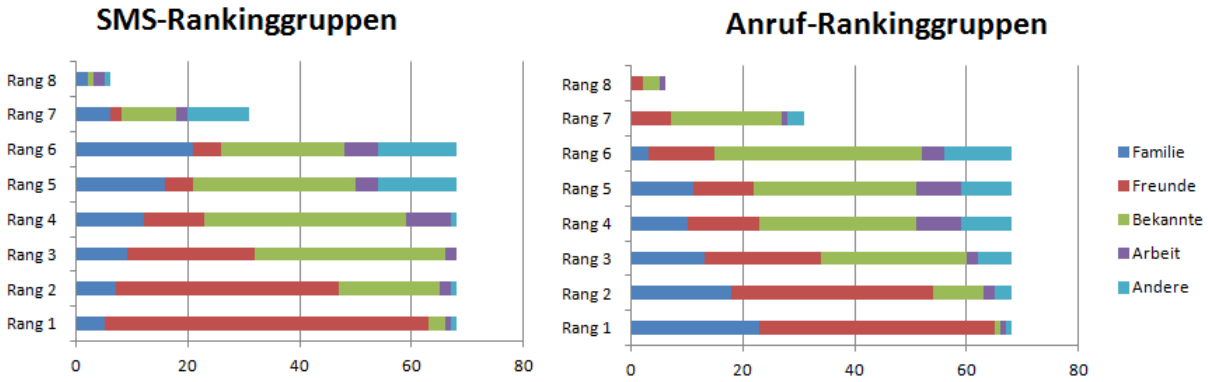


Abbildung 5.6.: Rankinggruppen für SMS- und Anruf-Aktivitäten. Auf der Abszissenachse werden die Anzahl an Kontakten dargestellt, deren Anteile durch die jeweiligen Farben gekennzeichnet sind. Auf der Ordinatenachse werden die Ränge dargestellt, in denen die Kontakte anhand ihrer SMS- und Anrufaktivität eingeordnet wurden.

6. Entwicklung von Klassifikatoren

Nachdem in den vorherigen Kapiteln die Sammlung und Auswertung der gesammelten Kontakt- und Aktivitätsdaten durchgeführt wurde, werden in diesem Kapitel nun Klassifikatoren entwickelt, welche Kontakte automatisch in eine der fünf vorgegebenen Gruppen einordnen können. Dazu werden zunächst die Informationen vorgestellt, welche direkt aus den gesammelten Kontakt- und Aktivitätsdaten auslesbar sind. Diese Informationen werden im Folgenden Variablen genannt. Anschließend werden Funktionen vorgestellt, die auf die Variablen angewendet werden können. Anhand der Variablen und den Funktionen werden daraufhin Attribute definiert, welche für die Entwicklung der Klassifikatoren eingesetzt werden. Die Begründung zur Verwendung eines Attributs wurde im vorherigen Kapitel im Diskussionsteil erläutert. Nach Entwicklung der Klassifikatoren werden diese im Anschluss evaluiert.

Es folgt die Aufzählung und Erläuterung der Variablen, welche direkt aus den Kontakt- und Aktivitätsdaten auslesbar sind.

Inhalt	**Variable**
Liste der an der Studie teilgenommenen Probanden	a
Liste aller Kontakte des Probanden p	k_p
Gibt an, ob Proband p neben SMS auch andere Nachrichtendienste verwendet.	u_p
Liste aller versendeten und empfangenen SMS Nachrichten des Probanden p.	sms_p
Liste aller getätigen Anrufe des Probanden p.	$calls_p$
Liste aller versendeten und empfangenen SMS Nachrichten des Kontakts c.	sms_c
Liste aller getätigten Anrufe des Kontakts c.	$calls_c$
Gibt die vom Probanden gewählte Kontaktgruppe für Kontakt c an	r_c

Tabelle 6.1.: Direkt aus den gesammelten Kontaktdaten auslesbare Variablen

6.1. Attribute

In diesem Unterkapitel werden Attribute anhand der Variablen in Tabelle 6.1 und den Funktionen in Tabelle 6.2 hergeleitet. Für jedes Attribut wird der Name, mit dem es im weiteren

Beschreibung	Funktion
Berechnet die durchschnittliche Länge aller gegebenen SMS-Nachrichten	*length()*
Berechnet die durchschnittliche Länge aller gegebenen Anrufe	*duration()*
Zählt die gegebenen SMS-Nachrichten bzw. Anrufe.	*count()*
Liefert den Timestamp der/des ältesten SMS/Anrufs	*first()*

Tabelle 6.2.: Auf die Variablen in Tabelle 6.1 anwendbaren Funktionen.

Verlauf eindeutig identifiziert werden soll, sowie die Herleitung und eine Beschreibung angegeben.

Gewichtete Aktivität Dieses Attribut gibt die Aktivität mit einem Kontakt c abhängig von der Nutzungshäufigkeit des Probanden p an. Dazu werden die SMS-Aktivität sowie die Anruf-Aktivität zunächst mit den jeweiligen Ausgleichsfaktoren a_{sms} und a_{calls} multipliziert und anschließend miteinander addiert. Je weniger SMS- beziehungsweise Anrufaktivität Proband p vorweist, desto größer sind diese Ausgleichsfaktoren. Dies bedeutet: je weniger der Proband p SMS-Nachrichten austauscht oder Anrufe tätigt, desto stärker wird eine einzelne SMS-Nachricht beziehungsweise ein einzelner Anruf an einen Kontakt c gewichtet und umgekehrt.

Berechnet werden die Ausgleichsfaktoren a_{sms} und a_{calls} mittels dem Verhältnis zwischen dem Timestamp der ersten Aktivität des Probanden p und dem Durchschnitt des ersten Timestamps aller Probanden. Das Ergebnis wird mit dem Verhältnis zwischen der Anzahl an Aktivitäten des Probanden p und der durchschnittlichen Anzahl an Aktivitäten von allen Probanden multipliziert. Zusätzlich wird der Ausgleichsfaktor a_{sms} mit einer Konstante u_p multipliziert. Dieser beträgt 0.75, wenn der Proband andere Nachrichtendienste neben den SMS-Nachrichten verwendet und 1.25, falls nicht.

Es folgt die Zusammensetzung der Variablen:

$$a_{sms} = \frac{first(sms_p)}{\frac{\sum_i^a first(sms_i)}{|a|}} \cdot \frac{count(sms_p)}{\frac{\sum_i^a count(sms_i)}{|a|}} \cdot u_p \tag{6.1}$$

$$a_{calls} = \frac{first(calls_p)}{\frac{\sum_i^a first(calls_i)}{|a|}} \cdot \frac{count(calls_p)}{\frac{\sum_i^a count(calls_i)}{|a|}} \tag{6.2}$$

Zusammen mit (6.1) und (6.2) berechnet folgende Formel das Attribut *weighted Activity*:

$$weighted\_activity = a_{sms} \cdot count(sms_c) + a_{calls} \cdot count(calls_c) \tag{6.3}$$

Verhältnis zwischen SMS und Anrufen Mit diesem Attribut wird das Verhältnis zwischen der Anzahl an SMS-Nachrichten und Anrufen repräsentiert. Dazu wird das prozentuale Vorkommen an SMS-Nachrichten in der gesamten Kommunikation mit einem Kontakt c angegeben. Dies wird wie folgt berechnet:

$$\frac{count(sms_c)}{count(sms_c) + count(calls_c)} \tag{6.4}$$

Anzahl an SMS-Nachrichten und Anrufe Diese Attribute geben die Anzahl an empfangenen sowie versendeten SMS-Nachrichten beziehungsweise die getätigten Anrufe in Bezug auf Kontakt c an. Der Ausdruck durch die definierten Variablen und Funktionen folgt:

$$count(sms_c) \tag{6.5}$$

und

$$count(calls_c) \tag{6.6}$$

Differenz zwischen lokaler und durchschnittlicher Anzahl an SMS und Telefonaten Hiermit wird die Differenz zwischen der durchschnittlichen Anzahl an SMS-Nachrichten bzw. Anrufen pro Kontakt von Proband p und der Anzahl an SMS-Nachrichten bzw. Anrufen für Kontakt c angegeben. Ist das Ergebnis positiv, so weist Kontakt c eine überdurchschnittliche Aktivität mit Proband p auf. Im negativen Fall bedeutet dies eine unterdurchschnittliche Aktivität.

$$\frac{count(sms_{all})}{|k_p|} - count(sms_c) \tag{6.7}$$

Ranggruppen für SMS-Nachrichten und Telefonaten Die Beschreibung sowie die Ermittlung von Ranggruppen wurde in Kapitel 5.3 erläutert.

Durchschnittliche SMS-Länge und Dauer eines Anrufs Diese Attribute geben die durchschnittliche Länge von SMS-Nachrichten beziehungsweise die durchschnittliche Dauer von Anrufen mit Kontakt c an. Berechnet werden diese wie folgt:

$$length(sms_c) \tag{6.8}$$

und

$$duration(calls_c) \tag{6.9}$$

Uhrzeiten Dieses Attribut gibt anhand einer Fließkommazahl (von 0..5) an, um welche Uhrzeiten die meiste Kommunikation mit diesem Kontakt stattfindet. Jede Zahl repräsentiert dabei einen 4-Stundenblock. Beispielsweise zeigt die Zahl 0 an, dass zwischen 0:00 und 3:59 Uhr die meiste Kommunikation stattfand, die Zahl 1.5 dagegen steht für den Block von 6:00 bis 9:59.

Für die Berechnung werden zunächst mithilfe der Timestamps alle Aktivitätszeiten auf einem Zeitstrahl markiert. Anschließend wird nach dem 4-Stundenblock mit den meisten Markierungen gesucht. Dazu iteriert man über alle 24 Stunden und schaut sich die aktuelle Stunde zusammen mit den 3 folgenden an.

Wochentage Dieses Attribut gibt an, ob die meisten Anrufe unter der Woche oder am Wochenende stattfanden. Um die Anzahl an Tage entsprechend zu gewichten, werden die Anzahl an ausgetauschten SMS-Nachrichten oder getätigten Anrufen unter der Woche durch 5 und die am Wochenende durch 2 geteilt.

Kontaktgruppe Dieses Attribut gibt die vom Probanden p angegebene Kontaktgruppe für den Kontakt c an. Nach diesem Attribut werden die entwickelten Klassifikatoren die Kontakte einordnen.

6.2. Entwicklung

Als Vorbereitung für die Entwicklung der Klassifikatoren wurde das in Kapitel 5 entwickelte *Hilfsprogramm* erweitert. Das *Hilfsprogramm* ist nun in der Lage die gesammelten Kontakt- und Aktivitätsdaten einzulesen und daraus mithilfe von WEKA automatisch einen Klassifikator zu entwickeln.

Zudem wurde, neben der bereits vorhandenen k-Fold-Cross-Validation von WEKA als Evaluierungsmethode, eine eigene Variante der Cross-Validation entwickelt. Statt der zufälligen Sortierung der Daten und anschließenden Aufteilung in k gleichgroße Folds [BFH+12, S. 16], wird ein komplettes Datenset eines einzelnen Probanden als ein Fold angesehen. Das heißt, dass hierbei die Daten von n-1 Probanden für den Trainingsprozess und das übrig gebliebene Datenset für den anschließenden Test verwendet werden.

Im Weiteren wird die k-Fold-Cross-Validation Variante von WEKA mit k=10 als *standard Cross-Validation* bezeichnet. Die Variante der Cross-Validation, bei der ein komplettes Datenset eines Probanden als Fold angesehen wird, wird als *Cross-Validation einzelner Probanden* bezeichnet.

Alle 439 Kontakte von 20 Probanden, welche in Kapitel 5 für die Auswertung verwendet wurden, werden hier in die Entwicklung der Klassifikatoren einbezogen.

Standard Cross-Validation				Cross-Validation einzelner Probanden			
Korrekt: 292 (66.5148 %)				Korrekt: 260 (59.2255 %)			
Inkorrekt: 147 (33.4852 %)				Inkorrekt: 179 (40.7745 %)			
Precision	Recall	F-Measure	Gruppe	Precision	Recall	F-Measure	Gruppe
0.521	0.494	0.507	Familie	0.523	0.541	0.479	Familie
0.752	0.806	0.778	Freunde	0.734	0.815	0.753	Freunde
0.661	0.765	0.709	Bekannte	0.597	0.493	0.480	Bekannte
0	0	0	Arbeit	0.146	0.093	0.104	Arbeit
0.694	0.581	0.633	Andere	0.296	0.291	0.241	Andere
0.618	0.649	0.632	(gew. Durchschn.)	0.620	0.571	0.594	(gew. Durchschn.)

Abbildung 6.1.: Erkennungsqualität des Standard J48 Klassifikators in der standard Cross-Validation und in der Cross-Validation einzelner Probanden.

6.2.1. Standard J48 Klassifikator

Für die Entwicklung des im Folgenden als *Standard J48 Klassifikator* benannten Klassifikators wurden alle in 6.1 vorgestellten Attribute verwendet. Als Lernmethode wurde J48 mit den Einstellungen (reducedErrorPruning=True) angewendet.

Werte zur Erkennungsqualität sind in Abbildung 6.1 abgebildet. Bei der standard Cross-Validation wurden von 439 Kontakten 292 korrekt klassifiziert, was einem Anteil von 66.5148% entspricht. Die gewichtete durchschnittliche Precision aller Gruppen liegt bei 61.8%, der Recall bei 64.9%. Beide kombiniert ergeben einen F-Measure von 63.2%. Bis auf die Kontaktgruppe *Arbeit* wurden Kontakte aller anderen Gruppen mit einem F-Measure von über 50% korrekt erkannt. Mit diesem Klassifikator konnte in der standard Cross-Validation kein Kontakt der Gruppe *Arbeit* korrekt erkannt werden.

Evaluiert man den Standard J48 Klassifikator mit einer Cross-Validation einzelner Probanden, so konnten 260 Kontakte von 439 korrekt klassifiziert werden. Die Precision liegt im gewichteten Durchschnitt bei 62.0% und der Recall bei 57.1%, was einen F-Measure von 59.44% ergibt. Im Vergleich zur standard Cross-Validation werden Kontakte der Gruppe *Andere* und *Bekannte* schlechter erkannt. Beide Gruppen kamen mit der Cross-Validation einzelner Probanden auf einen F-Measure von 24.1% beziehungsweise 48.0% im Vergleich zu 63.3% bzw. 70.9% in der standard Cross-Validation. Dagegen werden Kontakte der Gruppe *Arbeit* mit einem F-Measure von 10,4% erkannt. Die Erkennung der Kontaktgruppen *Familie* und *Freunde* sind um 2-3% schlechter.

Dabei liegt das schlechteste Ergebnis eines Folds bei der Cross-Validation einzelner Probanden bei 33.3% (17 von 51) an korrekt klassifizierten Kontakten. Es wurden keine Kontakte der Gruppe *Arbeit* und *Andere* korrekt erkannt. Ebenso gab es Probleme bei der Unterscheidung zwischen Kontakten der Gruppe *Familie* und *Bekannte*. Beim besten Ergebnis eines Folds wurden dagegen 30 von 37 Kontakten (=81.08%) korrekt klassifiziert. Nennenswert dabei ist, dass keine Kontakte der Gruppe *Arbeit* vorkamen. Die Werte Precision, Recall, F-Measure sowie die Confusion Matrix werden in Abbildung 6.2 dargestellt.

Bestes Ergebnis

```
Korrekt: 30 (81.0811 %)
Inkorrekt: 7 (18.9189 %)

Precision  Recall  F-Measure  Gruppe
  1.000    0.500    0.667    Familie
  0.750    0.900    0.818    Freunde
  0.947    0.818    0.878    Bekannte
  n/a      n/a      n/a      Arbeit
  0.250    1.000    0.400    Andere
  0.881    0.811    0.826    (gew. Durchschn.)

  a  b  c  d  e   <-- classified as
  2  0  1  0  1 |  a = Familie
  0  9  0  0  1 |  b = Freunde
  0  3 18  0  1 |  c = Bekannte
  0  0  0  0  0 |  d = Arbeit
  0  0  0  0  1 |  e = Andere
```

Schlechtestes Ergebnis

```
Korrekt: 17 (33.3333 %)
Inkorrekt: 34 (66.6667 %)

Precision  Recall  F-Measure  Gruppe
  0.268    0.541    0.407    Familie
  0.625    0.815    0.625    Freunde
  1.000    0.493    0.045    Bekannte
  0        0.093    0.104    Arbeit
  0        0.291    0.241    Andere
  0.598    0.333    0.239    (gew. Durchschn.)

  a  b  c  d  e   <-- classified as
 11  1  0  0  1 |  a = Familie
  3  5  0  0  0 |  b = Freunde
 20  1  1  0  0 |  c = Bekannte
  4  1  0  0  0 |  d = Arbeit
  3  0  0  0  0 |  e = Andere
```

Abbildung 6.2.: Bestes und schlechtestes Ergebnis für die Klassifizierung eines einzelnen Folds in der Cross-Validation einzelner Probanden.

6.2.2. Multilayer J48 Klassifikator

Die Idee des *Multilayer J48 Klassifikator* ist es, die vorgegebenen fünf Kontaktgruppen in drei neue Gruppen zusammenzufassen. Bei der Zusammenfassung entstehen somit zwei neue Kontaktgruppen bestehend aus jeweils zwei der ursprünglich vorgegebenen Gruppen (im Folgenden Obergruppe genannt) sowie einer Obergruppe, die aus einer vorgegeben Gruppe besteht. Anschließend werden insgesamt drei Klassifikatoren daraus trainiert: einer, der einen Kontakt in die 3 Obergruppen einordnet (im Folgenden Klassifikator A genannt) sowie zwei, welche Kontakte aus den zweiteiligen Obergruppen in ihre jeweilige vorgegebene Klasse kategorisieren (Klassifikator B und C). (Abbildung 6.3)

Um nun einen Kontakt einer Gruppe zuzuordnen, wird dieser zunächst mit Klassifikator A einer Obergruppe eingeordnet. Fällt der Kontakt auf eine der zweiteiligen Obergruppen, wird dieser anschließend mit dem entsprechenden Klassifikator B oder C in seine vorgegebene Gruppe klassifiziert.

Für die Entwicklung des Multilayer J48 Klassifikators wurden die Kontaktgruppen *Familie* mit *Andere* sowie *Freunde* mit *Bekannte* zusammengefasst, wobei die Gruppe Arbeit übrig blieb. Diese Entscheidung wurde zunächst durch das Verhältnis von SMS zu Anrufen (Abbildung 5.2) begründet. Um sicherzugehen, dass es keine besseren Kombinationen für die Zusammenfassung geben könnte, wurde das in Kapitel 5 entwickelte *Hilfsprogramm* so erweitert, dass alle Kombinationen systematisch ausprobiert und anschließend getestet wurden. [1]. Es stellte sich heraus, dass die genannte Wahl das beste Klassifizierungsergebnis für die Obergruppen bietet.

[1] Appendix A.4.2 zeigt, wie dieses Experiment selber durchgeführt werden kann

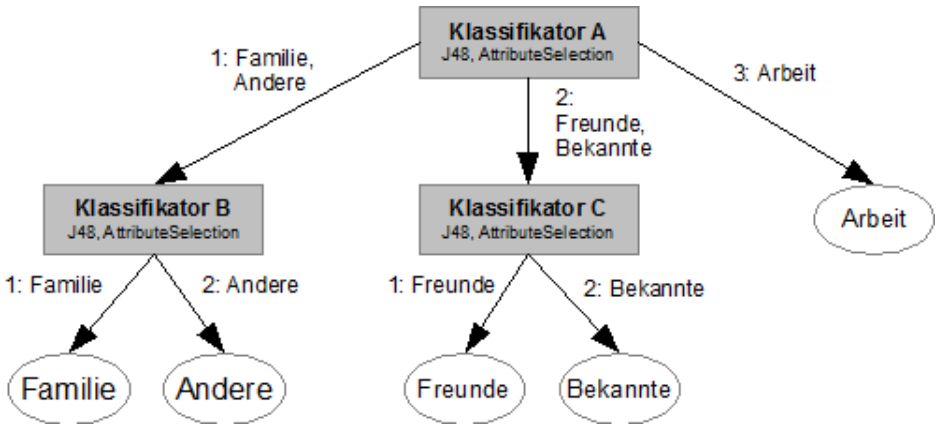


Abbildung 6.3.: Hintereinanderschaltung der Klassifikatoren A, B und C zum Multilayer J48 Klassifikator.

Bei einer standard Cross-Validation von Klassifikator A wurden von 439 Kontakten 342 (77.90%) richtig in ihre jeweiligen Obergruppen klassifiziert. Dabei wurde der *AttributeSelection* Filter mit den Optionen (evaluator=ClassifierSubsetEval(classifier=J48)) vor der Klassifizierung angewendet. Klassifikator B, der für die Unterscheidung von *Familie* und *Andere* zuständig ist, klassifizierte 98 von 120 Kontakten (81.67%) korrekt. Klassifikator C kommt bei der Unterscheidung von *Familienmitgliedern* und Kontakte der Gruppe *Andere* auf eine korrekte Klassifizierung von 257 von 292 Kontakten (88.01 %). Vor dem Trainieren von Klassifikator B und C wurde ebenfalls der *AttributeSelection* Filter mit den genannten Optionen verwendet.

Um den Multilayer J48 Klassifikator als Gesamtes zu testen, wurde in Java ein Programm entwickelt, welches Klassifikator A, B und C gemäß Abbildung 6.3 hintereinander schaltet. Bei der Cross-Validation einzelner Probanden klassifizierte dieser Klassifikator 256 von 439 Kontakten korrekt , was zu einem Anteil von 58.31% führt. Detailliertere Daten zur Erkennungsqualität des Multilayer J48 Klassifikators sind in Abbildung 6.4 abgebildet.

Der gewichtete Durchschnitt des F-Measure Wertes liegt bei 55.4%, wobei die Kontaktgruppen *Freunde* mit einem F-Measure von 73.4% und *Bekannte* mit einem F-Measure von 64.6% erkannt werden. Dagegen werden die Kontaktgruppen *Familie* und *Andere* lediglich mit F-Measure Werten von 39% und 28.6% erkannt. Keine Kontakte der Gruppe *Arbeit* wurden korrekt identifiziert.

Zu diesem Zeitpunkt konnte keine standard Cross-Validation durchgeführt werden, da WEKA in der Version 3.7.7 die Hintereinanderschaltung mehrerer Klassifikatoren zur Evaluierung nicht unterstützte.

6.2.3. Weitere Klassifikatoren

Mit den in Kapitel 6.1 genannten Attributen wurden weitere Lernmethoden für die Entwicklung eines Klassifikators verwendet. Da damit erstellte Klassifikatoren ihre Aufgabe

Cross-Validation einzelner Probanden

```
Korrekt: 256 (58.3144 %)
Inkorrekt: 183 (41.6856 %)

Precision  Recall  F-Measure  Gruppe
  0.390     0.390    0.390     Familie
  0.675     0.806    0.734     Freunde
  0.615     0.680    0.646     Bekannte
  0.0       0.0      0.0       Arbeit
  0.370     0.233    0.286     Andere
  0.533     0.595    0.554     (gew. Durchschn.)
```

Abbildung 6.4.: Erkennungsqualität des Multilayer J48 Klassifikator in der Cross-Validation einzelner Probanden.

Lernmethode	**Standard Cross-Validation**	**CV einzelner Probanden**
J48	66.5148% (292)	59.2255% (260)
ZeroR	34.8519% (153)	28.4738% (125)
RandomForest	66.0592% (290)	58.5421% (257)
SimpleCart	65.8314% (289)	57.4031% (252)
BFTree	63.5535% (279)	57.6309% (253)
Naive Bayes	63.3257% (278)	43.7357% (192)
Bayes Net	63.0979% (277)	60.13667% (264)
Functional Trees	61.0478% (268)	35.0797% (154)
DecisionTable	60.5923% (266)	54.6697% (240)
NBTree	60.3645% (265)	53.53075% (235)
RandomTree	60.3645% (265)	50.3416% (221)
Logistic Model Trees	59.6811% (262)	35.5353% (156)

Abbildung 6.5.: In dieser Tabelle werden mehrere Lernmethoden und die daraus resultierenden Klassifizierungsergebnisse in der standard Cross-Validation und der Cross-Validation einzelner Probanden aufgelistet. Ergebnisse werden dabei durch ihren Anteil sowie in Klammern durch die Anzahl an korrekt klassifizierten Kontakten dargestellt. Alle Lernmethoden wurden in ihrer Standardeinstellung verwendet.

nicht wesentlich effizienter erfüllen als die mit J48 entwickelten Klassifikatoren, werden die angewandten Lernmethoden anhand ihrer Ergebnisse in den zwei Varianten der Cross-Validation hier lediglich aufgezählt und nicht näher darauf eingegangen. Die Ergebnisse sind in Abbildung 6.5 abgebildet.

Standard Cross-Validation				Cross-Validation einzelner Probanden			
Korrekt: 153 (34.8519 %)				Korrekt: 125 (28.4738 %)			
Inkorrekt: 286 (65.1481 %)				Inkorrekt: 314 (71.5262 %)			
Precision	Recall	F-Measure	Gruppe	Precision	Recall	F-Measure	Gruppe
0	0	0	Familie	0	0	0	Familie
0	0	0	Freunde	0.032	0.154	0.052	Freunde
0.349	1	0.517	Bekannte	0.254	0.834	0.373	Bekannte
0	0	0	Arbeit	0	0	0	Arbeit
0	0	0	Andere	0	0	0	Andere
0.121	0.349	0.18	(gew. Durchschn.)	0.090	0.264	0.130	(gew. Durchschn.)

Abbildung 6.6.: Erkennungsqualität von ZeroR anhand der standard Cross-Validation und der Cross-Validation einzelner Probanden.

6.3. Evaluation

Mehrere Klassifikatoren mit der J48 Lernmethode wurden in diesem Kapitel entwickelt, welche nun mit den entsprechenden ZeroR Klassifikatoren verglichen sowie durch Probanden evaluiert werden.

6.3.1. Vergleich mit ZeroR

Die ZeroR Lernmethode bestimmt die am häufigsten vorkommende Kategorie in den Trainingsdaten und liefert dieses Ergebnis bei jeder Klassifizierung einer Instanz zurück. Dadurch wird ermittelt, wie gut Gruppen ohne Beachtung von Attributen vorhergesagt werden können. Ein Vergleich der bereits mit J48 entwickelten Klassifikatoren mit den ZeroR Vertretern soll die Verbesserung der Klassifizierung unter Beachtung der Attribute zeigen. [BFH+12, S. 16]

Beim standard Cross-Validation Verfahren wurde die Gruppe *Bekannte* als meistverwendete Gruppe identifiziert. Der daraus entstandene Klassifikator lieferte als Ergebnis somit stets die Gruppe *Bekannte* zurück. Im Gegensatz dazu gibt es bei der Cross-Validation einzelner Probanden einige Iterationen, bei denen Kontakte der Gruppe *Freunde* statt der Gruppe *Bekannte* am häufigsten in den Trainingsdaten vorkommen. Werte zur Erkennungsqualität der ZeroR Modelle werden in Abbildung 6.6 aufgelistet.

Ein Vergleich des Standard J48 Klassifikators (Abbildung 6.1) mit ZeroR durch die standard Cross-Validation zeigt eine Verbesserung von 31.6629% (139 Kontakte) auf. Dabei ist der Standard J48 Klassifikator in allen F-Measure Werten dem ZeroR Klassifikator überlegen. Gleiches ist bei einer Cross-Validation einzelner Probanden zu beobachten: Die Verbesserung des Standard J48 Klassifikators gegenüber dem ZeroR Klassifikator liegt bei 30,7517% (135 Kontakte). Der Multilayer J48 Klassifikator erzielte als Gesamtes eine Verbesserung von 29.16% gegenüber dem ZeroR Klassifikator in der Cross-Validation einzelner Probanden.

Trotz der Überlegenheit gegenüber dem ZeroR Klassifikator sowohl in der standard Cross-Validation als auch in der Cross-Validation einzelner Probanden, ist es möglich bei der Klassifizierung der Kontakte eines bestimmten Probanden ein schlechteres Ergebnis als der ZeroR Klassifikator zu erzielen. Als Beispiel dient das schlechteste Ergebnis in einem Fold der Cross-Validation einzelner Probanden für den Standard J48 Klassifikator. (s. Abbildung 6.2) Dieser Fall ist nicht vermeidbar, da es immer Anwender geben kann, die ein anderes Aktivitätsmuster gegenüber ihren Gruppen aufweisen als die große Mehrheit. In diesem Beispiel tätigte der verantwortliche Proband nahezu ausschließlich längere Anrufe mit seinen Kontakten der Gruppe *Bekannte*, weshalb diese vom Klassifikator als *Familienmitglieder* identifiziert wurden.

6.3.2. Evaluation durch Probanden

Um die Effizienz der Klassifikatoren in der Praxis zu evaluieren, wurde eine Android-Anwendung implementiert, welche Kontakte auf Smartphones anhand der beiden J48 Klassifikatoren gruppiert. An diesem Test, welcher drei Monate nach der Datensammlung (s. Kapitel 4 und 5) stattfand, nahmen fünf Probanden teil. Davon haben drei bereits an der Datensammlung teilgenommen, jedoch haben sich die Kontakt- und Aktivitätsdaten in den drei Monaten bei diesen stark verändert. Alle fünf Probanden waren Studenten und zwischen 19 und 24 Jahre alt.

Methodik Nach Installation der Anwendung wurden alle Kontakte, für die eine Aktivitäts-Historie existiert, ausgelesen und in die vorgegebenen fünf Gruppen klassifiziert. Anschließend konnte das Ergebnis in Form einer Liste angesehen werden. Dabei wurde die Möglichkeit geboten, falsch klassifizierte Kontakte zu markieren und diese manuell in die korrekte Gruppe einzuordnen. Alle Markierungen und manuelle Zuordnungen wurden protokolliert und bei Beenden der Anwendung in Form einer Textdatei gespeichert. Diese Protokolldaten werden im Folgenden ausgewertet.

Die Probanden bekamen mündlich die Anweisungen, zunächst das Ergebnis des Standard J48 Klassifikators zu untersuchen und anschließend falsche Klassifizierungen zu markieren, um eine andere Gruppe vorzuschlagen. Im Anschluss wurden diese Anweisungen ebenfalls für den Multilayer J48 Klassifikator durchgeführt.

Für die Auswertung der Studie wurden Confusion Matrices verwendet. Damit werden die korrekt klassifizierten Kontakte sowie die manuellen Korrekturen der Probanden visualisiert.

Ergebnisse Die Anwendung klassifizierte insgesamt 144 Kontakte, wobei am Ergebnis des Standard J48 Klassifikators insgesamt 51 Klassifizierungen als inkorrekt markiert wurden. Dagegen wurden 38 inkorrekte Klassifizierungen beim Multilayer J48 Klassifikator gemeldet. Das Ergebnis wurde in Abbildung 6.7 abgebildet.

Liest man die Confusion Matrix spaltenweise, so geben die Zellen (x, y), für die $x \neq y$ gilt, die Anzahl an Kontakten aus Gruppe y an, die fälschlicherweise in die Kontaktgruppe x klassifiziert wurden. Zieht man von der Summe der Spalte m den Wert der Zelle (m, m) ab, so erhält man die Anzahl an Korrekturen, die die Probanden von der Kontaktgruppe x in die Kontaktgruppe y korrigiert haben. Dieser Wert wird im Folgenden mit $\alpha_{<Kontaktgruppe>}$ angegeben.

Sowohl an der Confusion Matrix des Standard J48 Klassifikators als auch des Multilayer J48 Klassifikators ist anhand der α-Werte zu erkennen, dass die meisten Korrekturen an den fälschlicherweise in die Kontaktgruppe *Familie* eingeordneten Kontakten durchgeführt wurden. Die Anzahl an Korrekturen liegt hier bei insgesamt 23 für den Standard J48 Klassifikator und 15 für den Multilayer J48 Klassifikator. Am zweithäufigsten mit 11 beim Standard J48 Klassifikator und 12 beim Multilayer J48 Klassifikator wurden Kontakte markiert, die fälschlicherweise in die Kontaktgruppe *Arbeit* eingeordnet wurden. Dagegen wurden Kontakte, die inkorrekt in die Gruppen *Freunde* und *Bekannte* eingeordnet wurden, am seltensten in andere Gruppen verschoben. Für den Standard J48 Klassifikator liegt diese Anzahl für die Kontaktgruppe *Freunde* bei 3 und bei 5 für die Kontaktgruppe *Bekannte*. Der Multilayer J48 Klassifikator kategorisierte unzutreffenderweise 3 Kontakte als *Freunde* und 4 Kontakte als *Bekannte*, welche von Probanden markiert wurden.

Addiert man alle Zahlen in Zeile y miteinander, so gibt die Summe, welche im Folgenden mit β_y bezeichnet wird, an wie viele Kontakte den Probanden zu Folge zur Kontaktgruppe y gehören. Lediglich für die Kontaktgruppe *Familie* ergibt sich bei den zwei J48 Klassifikatoren die gleiche Summe für $\beta_{Familie}$. Für die Kontaktgruppen *Arbeit* und *Andere* liegen die Differenzen von β_{Arbeit} und β_{Andere} zwischen dem Standard und dem Multilayer J48 Klassifikator jeweils bei 1. Probanden gaben beim Standard J48 Klassifikator an, insgesamt 63 Kontakte der Gruppe *Freunde* und 55 Kontakte der Gruppe *Bekannte* zu besitzen. Dagegen wurde beim Multilayer J48 Klassifikator angegeben, 54 Kontakte in der Gruppe *Freunde* und 62 Kontakte in der Gruppe *Bekannte* seien korrekt. Die Differenz der Werte $\beta_{Freunde}$ und $\beta_{Bekannte}$ liegen hier somit bei jeweils 9 und 7.

Diskussion Aufgrund der hohen Werte $\alpha_{Familie}$ und α_{Arbeit} in jeweils beiden J48 Klassifikatoren lässt sich folgern, dass Kontakte, die fälschlicherweise als *Familie* oder *Arbeit* klassifiziert wurden, von Probanden am wahrscheinlichsten wahrgenommen werden. Dies bestätigen auch die Werte $\beta_{Familie}$ und β_{Arbeit}, welche die korrekte Anzahl an Kontakten in beiden Gruppen laut Probanden angeben. Beide Werte sind für beide genannten Gruppen nahezu gleich.

Im Gegensatz dazu sind die Werte $\alpha_{Freunde}$ und $\alpha_{Bekannte}$ relativ niedrig. Zudem sind auch die Werte $\beta_{Freunde}$ und $\beta_{Bekannte}$ mit je einer Differenz von 8 relativ verschieden. Übertragen auf die Studie bedeutet dies, dass Probanden sich eher unsicher sind, ob Kontakte jeweils in die Kontaktgruppen *Freunde* oder *Bekannte* gehören. Eine Unsicherheit in der Entscheidung bei Probanden bestätigen auch die Zellen $(3, 2)$ und $(2, 3)$ in beiden Confusion Matrices. Lediglich einmal bei den zwei Klassifikatoren wurden von den fünf Probanden ein Kontakt,

Standard J48 Klassifikator	Multilayer J48 Klassifikator
Korrekt: 93 (64.58 %) Inkorrekt: 51 (35.42 %)	Korrekt: 106 (73.61 %) Inkorrekt: 38 (26.39 %)
a b c d e <-- classified as 12 0 4 0 0 | a = Familie 12 37 1 9 4 | b = Freunde 10 0 38 2 5 | c = Bekannte 0 3 0 1 0 | d = Arbeit 1 0 0 0 5 | e = Andere	a b c d e <-- classified as 10 1 3 2 0 | a = Familie 7 43 0 4 0 | b = Freunde 6 0 47 6 3 | c = Bekannte 1 1 0 2 1 | d = Arbeit 1 1 1 0 4 | e = Andere

Abbildung 6.7.: Confusion Matrix der Evaluierung durch Probanden. Diese geben neben der Anzahl an korrekt klassifizierten Kontakten an, wie viele Kontakte von Probanden manuell in ihre korrekte Kontaktgruppe korrigiert wurden.

welcher als *Freund* klassifiziert wurde, manuell in die Kontaktgruppe *Bekannte* verschoben. Umgekehrt ist dieser Fall nicht aufgetreten.

6.4. Fazit

In diesem Kapitel wurden zunächst Attribute hergeleitet, anhand denen mehrere Klassifikatoren entwickelt wurden. Diese Klassifikatoren wurden anschließend durch verschiedene Methoden evaluiert.

In der *standard Cross-Validation* erreichte der Standard J48 Klassifikator mit 292 von 439 korrekt klassifizierten Kontakten das beste Ergebnis. Ebenso erreichte der Standard J48 Klassifikator in der *Cross-Validation einzelner Probanden* das beste Ergebnis mit 59.2 % an korrekt klassifizierten Kontakten. Der Multilayer J48 Klassifikator konnte dagegen 58.3 % aller Kontakte korrekt klassifizieren.

Eine Betrachtung der *Evaluation durch Probanden* zeigt jedoch, dass am Ergebnis des Multilayer J48 Klassifikators weniger inkorrekte Klassifizierungen gefunden wurden als am Ergebnis des Standard J48 Klassifikators. Dabei konnte beobachtet werden, dass es Probanden kaum stört, wenn fälschlicherweise ein Kontakt der Gruppe *Freunde* zur Kontaktgruppe *Bekannte* oder umgekehrt eingeordnet wurde. Auf der anderen Seite bemerkten es Probanden allerdings sofort, wenn Arbeitskollegen oder Familienmitglieder mit anderen Kontaktgruppen verwechselt wurden. Aufgrund dieses Verhaltens konnte der Multilayer J48 Klassifikator durch seine Hintereinanderschaltung von 3 Klassifikatoren ein besseres Ergebnis erzielen als der Standard J48 Klassifikator.

Ein Vergleich der vorgestellten Klassifikatoren mit denen, die durch die ZeroR Lernmethode entwickelt wurden, zeigt eine Verbesserung von durchschnittlich 30% an korrekt klassifizierten Kontakten. Trotzdem kann eine Klassifikation auf bestimmten Datensets schlechter abschneiden als ein ZeroR Klassifikator. Dies wäre dann der Fall, wenn das Aktivitätsmuster eines Anwenders sehr verschieden ist zu dem der Probanden, welche die Trainingsdaten zur Verfügung stellen.

Die Effizienz der Klassifikatoren hängen somit hauptsächlich von den Trainingsdaten und der dazu passenden Zielgruppe ab. Stimmen diese nicht überein, so können dies weder die Attribute noch die Lernmethoden ausgleichen. Aus diesem Grund wäre es sinnvoll, für verschiedene Zielgruppen (beispielsweise Studenten oder Berufstätige) verschiedene Klassifikatoren basierend auf passenden Trainingsdaten zu entwickeln.

7. Fazit und Ausblick

In dieser Arbeit wurden verschiedene Klassifikatoren entwickelt, welche Kontakte automatisch in eine dafür vorgesehene Kontaktgruppe einordnen. Dazu wurde zunächst im Rahmen einer Studie ermittelt, welche Kontaktgruppen von Probanden bevorzugt zur Kategorisierung ihrer Kontakte verwendet wurden. Darauf aufbauend wurde ein System entwickelt, das Kontakt- und Aktivitätsdaten aus Smartphones ausliest. Die dabei ausgelesenen Kontakte wurden anschließend manuell von Probanden in eine der fünf bevorzugten Kontaktgruppen eingeordnet und für die weitere Arbeit gesammelt. Diese Daten wurden später ausgewertet, um nach auffälligen Verhaltensmustern in den einzelnen Gruppen zu forschen. Das daraus gewonnene Wissen wurde letztendlich für die Entwicklung von Klassifikatoren eingesetzt.

Bei der Auswertung der Kontakt- und Aktivitätsdaten ging es darum Merkmale von Kontaktgruppen ausfindig zu machen, anhand denen sich eine Kontaktgruppe von den restlichen unterscheiden lässt. Unter diesen Merkmalen sind unter anderem die Verhältnisse zwischen SMS- und Anrufaktivitäten, die in Kapitel 5.1 erwähnten Rankinggruppen für SMS- und Anrufaktivitäten sowie die durchschnittliche Dauer von Telefonaten zu erwähnen. Anhand dieser Merkmale, welche in Kapitel 6 zu Attributen für die Klassifizierung umgeschrieben wurden, wurden anschließend verschiedene Klassifikatoren entwickelt.

Es wurden insgesamt zwei verschiedene Klassifikatoren anhand der J48 Lernmethode sowie 10 weitere Klassifikatoren mithilfe anderer Lernmethoden entwickelt. Nach der jeweiligen Entwicklung eines Klassifikators wurde dieser mit zwei Varianten der Kreuzvalidierung evaluiert. Am Schluss wurden die zwei anhand der J48 Lernmethode entwickelten Klassifikatoren in eine Android-Anwendung eingebunden und anhand von Probanden evaluiert. Zudem wurden die entwickelten Klassifikatoren mit den Klassifikatoren verglichen, welche mit der ZeroR Lernmethode entwickelt wurden. Auf diese Weise konnte die Effizienz der gewählten Attribute herausgefunden werden.

Das beste Ergebnis in der standard Cross-Validation erreichte der Standard J48 Klassifikator mit 292 von 439 korrekt klassifizierten Kontakten, was einen korrekten Anteil von 66.5 % ausmacht. Dabei wurden Kontakte aus vier von fünf Gruppen mit einem F-Measure von über 50 % korrekt erkannt. Die Kontaktgruppen *Freunde* und *Bekannte* wurden dabei mit einem F-Measure von 77 % und 70 % korrekt identifiziert. Dagegen konnten Kontakte der Kontaktgruppe *Arbeit* nicht korrekt erkannt werden, was zu einem F-Measure von 0 % führte. Ein Vergleich mit den Klassifikatoren, welche mit der ZeroR Lernmethode entwickelt wurden, zeigte eine Verbesserung von durchschnittlich 30 % auf. In der anschließenden Evaluation durch Probanden wurden Ergebnisse in Höhe von 64.58 % für den Standard J48 Klassifikator und 73.61 % für den Multilayer J48 Klassifikator erreicht. Dabei wurden Confusion Matrices für die falsch und korrekt klassifizierten Kontakte der Probanden aufgestellt. Die Diskussion

zeigte unter anderem, dass Probanden nur ungenau zwischen den Kontaktgruppen *Freunde* und *Bekannte* unterscheiden konnten. Dagegen wurden jedoch falsche Klassifizierungen der Kontakte aus den Gruppen *Familie* und *Arbeit* sehr schnell erkannt.

7.1. Fazit

Ziel der vorliegenden Arbeit war es einen Klassifikator zu entwickeln, welcher die Gruppierung der Kontakte auf Smartphones automatisch durchführt. Zu diesem Zweck wurden zunächst die manuellen Gruppierungen der Anwender analysiert und anschließend darauf aufbauend verschiedene Klassifikatoren entwickelt. Die Tests dieser Klassifikatoren durch zwei Varianten der Cross-Validation zeigten, dass durchschnittlich 60 % aller Kontakte korrekt in ihre Kontaktgruppen eingeordnet werden konnten. Aus einer anschließenden Evaluation durch Probanden folgte, dass diese mit bis zu 73% aller Klassifizierungen zufrieden waren. Im Gegensatz zu formalen Testverfahren, wie beispielsweise der Cross-Validation, gibt es bei Anwendern keine strikte Unterscheidung zwischen einigen Kontaktgruppen. So zeigte die Evaluation durch Probanden, dass sich Anwender bei vielen Kontakten nicht sicher waren, ob sie diese in die Kontaktgruppe *Freunde* oder in die der *Bekannten* einordnen würden.

Da hauptsächlich Studenten für die Studien dieser Arbeit rekrutiert wurden, wurden die Klassifikatoren dementsprechend auch nur mit Klassifizierungen durch studentische Probanden trainiert. Dies begründet auch den niedrigen Anteil von 6.5 % an Kontakten der Gruppe *Arbeit* in den gesammelten Kontakt- und Aktivitätsdaten und somit die entsprechenden F-Measure-Werte für die Kontaktgruppe *Arbeit*. Daraus kann geschlossen werden, dass die Güte der entwickelten Klassifikatoren von den gesammelten Kontakt- und Aktivitätsdaten sowie der Zielgruppe, für die die Kontakte klassifiziert werden sollen, abhängig ist. So werden Klassifikatoren, die anhand von Daten studentischer Probanden entwickelt wurden, nur wenige Kontakte auf Smartphones von Beschäftigten korrekt klassifizieren können.

Ein Vergleich der entwickelten Klassifikatoren mit den Ergebnissen aus verwandten Arbeiten (beispielsweise wurden in [GK09] eine Genauigkeit von 85% bei der Berechnung der tie strength erreicht) zeigt auf den ersten Blick einen signifikanten Unterschied in der Güte der Gruppierung. Eine nähere Betrachtung der zur Verfügung stehenden Daten jedoch zeigt, dass in den genannten Artikeln durch soziale Netzwerke oder gemeinsame Empfänger in E-Mail-Nachrichten mehr Informationen für die Klassifizierung verfügbar waren. Diese Menge an Informationen ist heutzutage auf Smartphones nicht verfügbar, sodass hier lediglich auf den SMS- und Anrufsverlauf zurückgeblickt werden konnte. Zudem wurden in den verwandten Arbeiten keine vorgegebenen Gruppen für die Klassifizierung verwendet. Ergebnisse in der Gruppierung von Kontakten auf Smartphones können somit nicht mit denen aus der Klassifizierung von Kontakten in sozialen Netzwerken verglichen werden.

Abschließend kann festgestellt werden, dass den Smartphone-Anwendern einiges an Aufwand durch den Einsatz der entwickelten Klassifikatoren abgenommen werden konnten.

Dies wurde bei der Evaluation durch Probanden mit einem Ergebnis von 73% an korrekt klassifizierten Kontakten bestätigt. Jedoch ist in den meisten Fällen anschließend eine manuelle Korrektur an kleinen Teilen der Klassifikationen nötig.

7.2. Ausblick und weitere Arbeit

Die Ergebnisse der Klassifikatoren in der Evaluation zeigten, dass die Erkennung von Kontakten der Gruppe *Arbeit* noch nicht zufriedenstellend ist. Dies liegt hauptsächlich an den gesammelten Kontakt- und Aktivitätsdaten, welche alle von studentischen Probanden stammen, die noch nicht arbeiteten und somit keine Arbeitskollegen in ihrer Kontaktliste besaßen.

Es liegt somit nahe für verschiedene Zielgruppen (beispielsweise Studenten, Beschäftigte oder Rentner) einerseits verschiedene Probanden der jeweiligen Zielgruppe zu rekrutieren und andererseits auch verschiedene Kontaktgruppen zur Klassifizierung zu verwenden. Es wäre beispielsweise sinnvoll bei der Klassifizierung auf Smartphones von Studenten die Kontaktgruppe *Arbeit* zu entfernen und diese durch eine andere, sinnvollere Kontaktgruppe zu ersetzen. Auf diesen Gedanken aufbauend, können verschiedene Klassifikatoren für verschiedene Zielgruppen entwickelt werden. Vor einer Klassifizierung der Kontakte werden Anwender dann aufgefordert ihre Zielgruppe auszuwählen, um so den entsprechenden Klassifikator einsetzen zu können.

Neben einer Klassifikation, die hauptsächlich auf der Menge an Anrufen oder SMS-Nachrichten basiert, wäre zusätzlich eine Klassifikation mithilfe des Inhalts von SMS-Nachrichten denkbar. So kann ähnlich wie in [GK09] eine SMS-Nachricht nach bestimmten Wörtern durchsucht werden, die in den meisten Fällen nur an bestimmte Kontaktgruppen gesendet werden. Gilbert et al. nannte im erwähnten Artikel Wörter wie “Geburtstag”, “Schatz” oder “beschäftigt” als Beispiel.

Die Einsatzgebiete solcher Klassifikatoren wären einerseits die Erleichterung der Suche nach Kontakten, andererseits wären auch Verwendungen in automatisierenden Anwendungen wie beispielsweise *Tasker* [1]denkbar. So kann damit ein dynamisches Telefonbuch konzipiert werden, welches zu einer bestimmten Uhrzeit und einem bestimmten Wochentag eine dafür geeignete Kontaktgruppe besonders hervorhebt und den Zugriff auf diese erleichtert. Beispielsweise werden Arbeitskollegen zu Arbeitszeiten an den ersten Stellen des Telefonbuchs angezeigt, während am Wochenende die Familienmitglieder oder die Freunde hervorgehoben werden. Auch kann die eigene Kontaktierbarkeit davon abhängig gemacht werden, in welcher Gruppe sich ein Kontakt befindet. Während zu Arbeitszeiten alle Anrufe von Arbeitskollegen oberste Priorität haben, werden Anrufe von Bekannten nur durch Vibration gekennzeichnet oder automatisch abgewiesen.

[1]Tasker ist eine Automatisierungsanwendung, mit der Aufgaben basierend auf einem Kontext (Zeit, Ort oder Event) durchgeführt werden können. Homepage: http://tasker.dinglisch.net/ *(zuletzt geprüft: 28.11.2012)*

Im Zeitalter von Smartphones ist es nicht unüblich private Informationen wie den aktuellen Standort oder spontane Schnappschüsse durch soziale Anwendungen mit anderen Kontakten zu teilen. Auch hier ist ein Einsatz der entwickelten Klassifikatoren denkbar. Durch die Gruppierung der Kontakte kann nun festgelegt werden, mit welcher Kontaktgruppe welche Informationen geteilt werden sollen. So sollen bestimmte Bilder nur mit Freunden oder Familienmitgliedern geteilt werden, jedoch nicht mit den Arbeitskollegen oder den Bekannten.

A. Appendix

A.1. Beispiel einer ARFF Datei

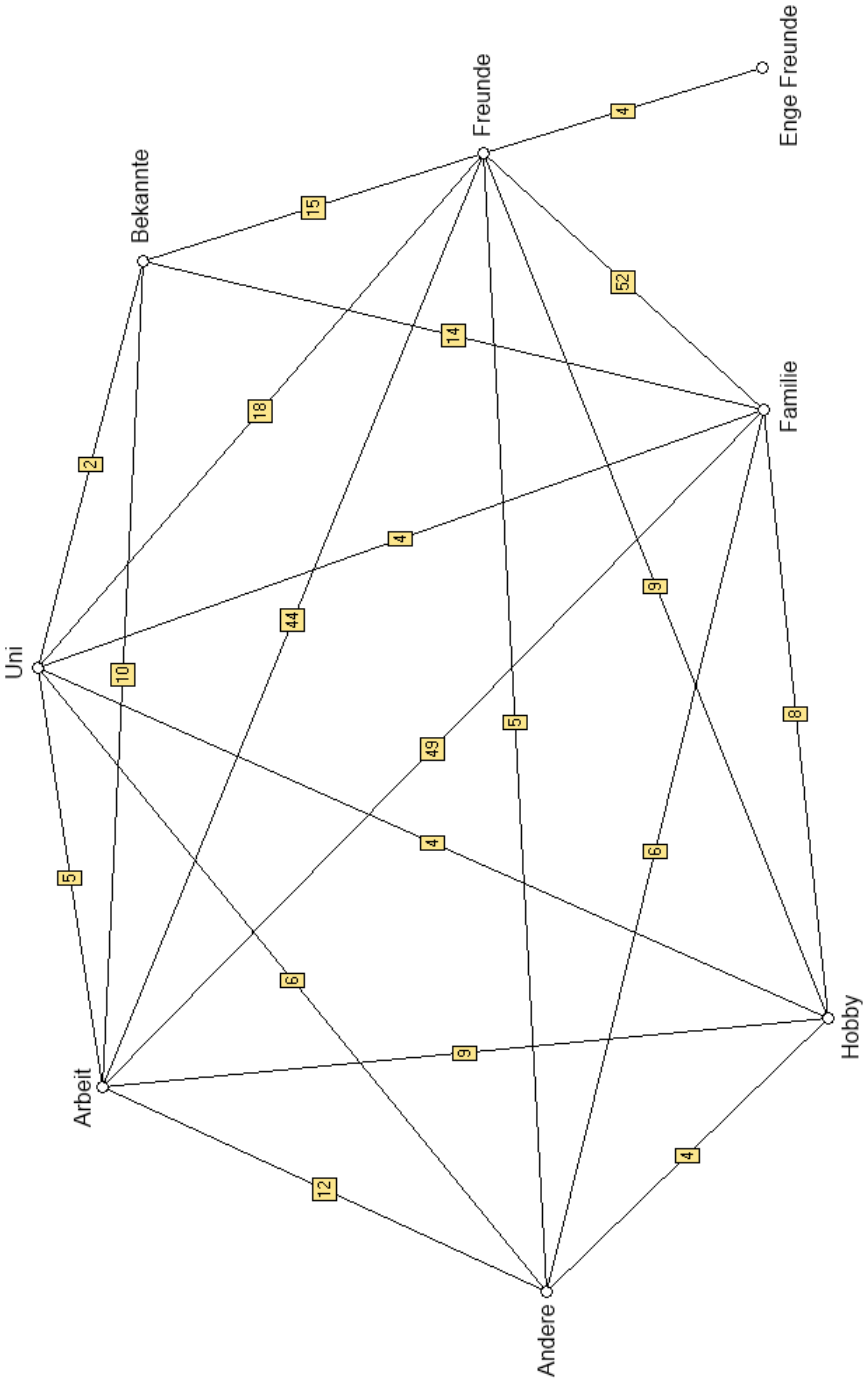
```
@relation weather
@attribute Aussicht {sonnig, bewölkt, regnerisch}
@attribute Feuchtigkeit {hoch, normal}
@attribute Wind {ja, nein}
@attribute Spielen {ja, nein}

@data
sonnig,hoch,nein,nein
sonnig,hoch,ja,nein
bewölkt,hoch,nein,ja
regnerisch,hoch,nein,ja
regnerisch,normal,nein,ja
regnerisch,normal,ja,nein
bewölkt,normal,ja,ja
sonnig,hoch,nein,nein
sonnig,normal,nein,ja
regnerisch,normal,nein,ja
sonnig,normal,ja,ja
bewölkt,hoch,ja,ja
bewölkt,normal,nein,ja
regnerisch,hoch,ja,nein
```

Eine ARFF-Datei besteht aus einem Header und dem Datenteil. Im Header werden Attribute definiert, wobei ein Attribut entweder ein nominal oder oder numerisch sein kann. Unter dem Header folgt der Datenteil. Instanzwerte der Attribute werden durch Kommata getrennt angegeben. Die Reihenfolge der Werte entspricht der Reihenfolge der Deklaration der Attribute im Header.

Nominale Attribute erfordern hinter ihrem Namen die Nennung aller möglichen Nominalwerte. Bei numerischen Attributen sind hinter dem Namen keine Angaben zu machen.

A.2. Visualisierung gemeinsam vorkommender Kontaktgruppen


Enge Freunde
Freunde
Bekannte
Familie
Uni
Arbeit
Hobby
Andere
4
15
52
14
18
2
4
9
44
10
5
8
49
4
6
5
6
9
4
12

A.3. Fragebogen

Geschlecht

- männlich
- weiblich

Alter

- < 14
- 14-17
- 18-25
- 25-50
- > 50

Frage 1: Wie viele Kontakte haben Sie ungefähr in Ihrer Kontaktliste ihres Mobiltelefons?

- < 50
- >50 & <100
- >100 & <200
- >200 & <400
- >400

Frage 2: Wie oft verwenden Sie Ihr Mobiltelefon zum Telefonieren?

- Mehrmals täglich
- Täglich
- Wöchentlich
- Monatlich
- Nie

Frage 3: Wie oft verwenden Sie Ihr Mobiltelefon um SMS Nachrichten zu versenden?

- Mehrmals täglich
- Täglich
- Wöchentlich
- Monatlich
- Nie

Frage 4: Angenommen Sie müssten Ihre Kontakte in verschiedene Gruppen einordnen. Welche Gruppennamen würden Sie verwenden?

A.4. Verwendung des Hilfsprogramms

In diesem Abschnitt soll gezeigt werden, wie das Hilfsprogramm verwendet oder erweitert werden kann. Der Quellcode des Hilfsprogramms ist in folgende 6 Pakete aufgeteilt:

- Main
  Dieses Paket beinhaltet die Hauptklasse, welche für das Einlesen sowie das Konvertieren von JSON-Dateien in ein *Proband*-Objekt zuständig ist.
- Analyser
  Dieses Paket beinhaltet alle Klassen, die für die Verarbeitung der *Proband*-Objekte zuständig sind. Dabei werden die einzelnen Kontakte und ihre Daten, zu denen die SMS-Nachrichten und Telefonate in Form von Länge/Dauer und Zeitpunkt, zu den in Kapitel 6.1 vorgestellten Attributen verarbeitet.
- json
  Enthält die Klassen, die für den Import einer JSON-Datei benötigt werden. Die JSON-Dateien werden zu Objekte dieser Klassen deserialisiert.
- tools
  Beinhaltet eine Klasse mit Funktionen, die die Arbeit in den Klassen des Pakets *Analyser* erleichtert. Zu diesen gehört beispielsweise die Berechnung des Durchschnitts oder der Standardabweichung.
- classifiergen
  Die in diesem Paket enthaltenen Klassen implementieren die Cross-Validation einzelner Probanden und erstellen Klassifikatoren mithilfe von WEKA und anhand der *Proband*-Objekte.
- res
  In diesem Paket sind die Header von ARFF-Dateien gespeichert.

Zudem wurden die Bibliotheken Jackson [1] und WEKA eingebunden. Der Pfad zu den JSON-Dateien wird in der Main-Klasse über der main-Methode eingestellt. Die Variablen, die der Einstellung dienen, wurden dabei durch Kommentare kenntlich gemacht.

A.4.1. Konvertierung ins ARFF-Format und Einführung neuer Attribute

Das Einlesen von JSON-Dateien und die anschließende Deserialisierung zu einem *Proband*-Objekt geschieht in der Main-Klasse. In den Klassen des Analyser-Pakets wurden die Klassen implementiert, an die jeweils ein Proband-Objekt übergeben werden. Diese Proband-Objekte werden in den genannten Klassen, welche alle die abstrakte Klasse *AbstractAnalyser* erweitern, analysiert und letztendlich daraus Attribute gebildet. Dabei entspricht der *output*-Vector eines Probanden den Instanzen einer ARFF-Datei.

[1] *Jackson High-performance JSON processor*: http://jackson.codehaus.org/ *(zuletzt geprüft: 27.11.2012)*

A.4.2. Suche nach besten Obergruppen für den Multilayer J48 Klassifikator

Die Suche nach der besten Obergruppe für den Multilayer J48 Klassifikator wurde in der Klasse `FIND_BEST_MLJ48CLASSIFIER` im Paket `classifiergen` implementiert. Um diese Klasse zu verwenden, muss in der Main-Klasse in der Methode `executeWeka()` ein Objekt der genannten Klasse erzeugt werden. Die Daten werden direkt aus den Attempt3-Klassen importiert.

A.5. Inhalt der CD-ROM

Auf der beiliegenden CD-ROM sind folgende Inhalte enthalten:

- Der Standard J48 Klassifikator und der Multilayer J48 Klassifikator als .model-Dateien für WEKA und als Java-Quellcode.
- Quellcode des *Hilfsprogramms*
- Quellcode des Systems zur Sammlung von Kontakt- und Aktivitätsdaten. Dazu gehört der Quellcode der Android-Anwendung, das PHP-Skript zum Speichern der Daten in die Datenbank sowie ein SQL-Code, der die benötigten Tabellen in der Datenbank erstellt.
- Quellcode der Demo-Anwendung für den Standard J48 Klassifikator und Multilayer J48 Klassifikator, mit der die Evaluation in Kapitel 6 durchgeführt wurde.
- Die vorliegende Arbeit in Form einer PDF- sowie Latex-Datei

Literaturverzeichnis

[AFW12] S. Amershi, J. Fogarty, D. Weld. Regroup: interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, S. 21–30. ACM, New York, NY, USA, 2012. (Zitiert auf Seite 12)

[ASL+09] A. Ankolekar, G. Szabo, Y. Luon, B. A. Huberman, D. Wilkinson, F. Wu. Friendlee: a mobile application for your social life. In *Proceedings of the 11th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '09, S. 27:1–27:4. ACM, New York, NY, USA, 2009. (Zitiert auf Seite 11)

[BFH+12] R. R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, D. Scuse. WEKA Manual for Version 3-7-7, 2012. (Zitiert auf den Seiten 48 und 53)

[GK09] E. Gilbert, K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, S. 211–220. ACM, New York, NY, USA, 2009. (Zitiert auf den Seiten 12, 60 und 61)

[Gra73] M. S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, S. 1360–1380, 1973. (Zitiert auf Seite 12)

[KE04] A. Kemper, A. Eickler. *Datenbanksysteme - Eine Einführung, 5. Auflage*. Oldenbourg, 2004. (Zitiert auf Seite 30)

[KL09] A. Komninos, D. Liarokapis. The use of mobile contact list applications and a context-oriented framework to support their design. In *Mobile HCI*, ACM International Conference Proceeding Series. ACM, 2009. (Zitiert auf Seite 11)

[Koh95] R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *IJCAI*, S. 1137–1145. 1995. (Zitiert auf Seite 17)

[MHT+11] D. MacLean, S. Hangal, S. K. Teh, M. S. Lam, J. Heer. Groups without tears: mining social topologies from email. In *Proceedings of the 16th international conference on Intelligent user interfaces*, IUI '11, S. 83–92. ACM, New York, NY, USA, 2011. (Zitiert auf Seite 12)

[MRS08] C. Manning, P. Raghavan, H. Schütze. *Introduction to information retrieval*, Band 1. Cambridge University Press Cambridge, 2008. (Zitiert auf Seite 17)

[ORT05] A. Oulasvirta, M. Raento, S. Tiitta. ContextContacts: re-designing SmartPhone's contact book to support mobile awareness and collaboration. In *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, MobileHCI '05, S. 167–174. ACM, New York, NY, USA, 2005. (Zitiert auf Seite 11)

[PMT+11] T. J. Purtell, D. MacLean, S. K. Teh, S. Hangal, M. S. Lam, J. Heer. An Algorithm and Analysis of Social Topologies from Email and Photo Tags. In *Workshop on Social Network Mining & Analysis, ACM KDD*. 2011. (Zitiert auf Seite 12)

[POL+09] A.-K. Pietiläinen, E. Oliver, J. LeBrun, G. Varghese, C. Diot. MobiClique: middleware for mobile social networking. In *Proceedings of the 2nd ACM workshop on Online social networks*, WOSN '09, S. 49–54. ACM, New York, NY, USA, 2009. (Zitiert auf Seite 11)

[Qui93] J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993. (Zitiert auf Seite 16)

[RBDD+10] M. Roth, A. Ben-David, D. Deutscher, G. Flysher, I. Horn, A. Leichtberg, N. Leiser, Y. Matias, R. Merom. Suggesting friends using the implicit social graph. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, S. 233–242. ACM, New York, NY, USA, 2010. (Zitiert auf Seite 12)

[Run10] T. A. Runkler. Data Mining - Methoden und Algorithmen intelligenter Datenanalyse, 2010. (Zitiert auf Seite 16)

[SPKG12] V. Stefanis, A. Plessas, A. Komninos, J. Garofalakis. Patterns of usage and context in interaction with communication support applications in mobile devices. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, MobileHCI '12, S. 25–34. ACM, New York, NY, USA, 2012. (Zitiert auf Seite 11)

[VRRK11] S. Visa, B. Ramsay, A. L. Ralescu, E. van der Knaap. Confusion Matrix-based Feature Selection. S. 120–127. 2011. (Zitiert auf Seite 18)

[WF05] I. H. Witten, E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2nd Auflage, 2005. (Zitiert auf Seite 16)

[WFT+99] I. H. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, S. J. Cunningham. Weka: Practical Machine Learning Tools and Techniques with Java Implementations, 1999. (Zitiert auf Seite 18)

Erklärung

Hiermit versichere ich, diese Arbeit selbständig verfasst und nur die angegebenen Quellen benutzt zu haben.

(Huy Viet Le)