Eberhard Karls Universität Tübingen
Mathematisch-Naturwissenschaftliche Fakultät
Wilhelm-Schickard-Institut für Informatik

Master Thesis Computer Science

Tangles on Ordinal Data

Alexander Conzelmann

Datum

Reviewers

Prof. Dr. Ulrike von Luxburg
Theory of Machine Learning
Wilhelm-Schickard-Institut für Informatik
Universität Tübingen

Prof. Felix Wichmann, DPhil
Neural Information Processing
Wilhelm-Schickard-Institut für Informatik
Universität Tübingen

Conzelmann, Alexander
Tangles on Ordinal Data
Master Thesis Computer Science
Eberhard Karls Universität Tübingen
Thesis period: 04/2022-10/2022

Abstract

Write here your abstract.

Zusammenfassung

Bei einer englischen Masterarbeit muss zusätzlich eine deutsche Zusammenfassung verfasst werden.

Acknowledgements

Write here your acknowledgements.

Contents

List of Figures

List of Tables

List of Abbreviations

BLAST	Basic Local Alignment Search Tool
...	...

Chapter 1

Introduction

Start with a comprehensive introduction about the questions of your thesis.

The thesis could include a background section, which also could become one or two separate chapters.

Do not forget to also give a short overview of the structure of the thesis in this chapter, for example as follows:

This thesis is structured as follows: First a background on XXX is introduced in the following background chapter (or the following section). In chapter **??** the developed algorithm to analyse ... is presented, followed by a comprehensive description of the used data or material. The results are given in chapter **??**. A discussion and short outlook conclude this thesis.

The following (in german) help newbies in LaTeXto learn about sections, math equations and much more.

Bevor wir uns der Auswertung bzw. Bewertung der gewonnenen Primärdaten zuwenden, wollen wir zunächst einige grundlegende Begriffe der deskriptiven Statistik wiederholen.

1.1 Stichproben

Grundsätzlich haben wir es bei Microarrayexpressionsdaten mit einer *Stichprobe* aus einer *Population (Grundgesamtheit)* zu tun. Wir bezeichnen nun im allgemeinen mit $X = \{x_1, x_2, \ldots, x_n\}$ die Beobachtungsdaten vom Umfang n. Diese Daten sollen mit statistischen Kenngrößen beschrieben werden. Aus diesen will man möglichst zuverlässig auf die zugrundeliegende Verteilung in der Grundgesamtheit schließen. Hierzu verwenden wir die **Lage-** und **Streuparameter**. Zunächst wenden wir uns aber der Häufigkeits- und Summenhäufigkeitsverteilung zu, die sowohl graphisch als auch numerisch einen Eindruck über die Verteilung von X bieten. Dafür betrachten wir diskrete Verteilungen.

Gegeben sei eine Stichprobe $(X_1, X_2, \ldots, X_n)$. Eine Funktion $Z_n = Z(X_1, \ldots, X_n)$ heisst eine *Stichprobenfunktion.* Sie ist selber eine Zufallsgröse.

1.1.1 Häufigkeiten und Histogramm

In X trete der Wert x_i genau n_i mal auf, $i = 1, 2, \ldots m$. Dann ist $\sum_i n_i = n$. Der Quotient n_i/n ist die *relative Häufigkeit* für das Eintreten des Ereignisses "$X = x_i$". Die Menge der relativen Häufigkeiten $\{n_1/n, n_2/n, \ldots, n_m/n\}$ heiśt *Häufigkeitsverteilung* von X. Ferner heiśt die Menge $\{s_1, \ldots, s_m\}$ mit $s_i = \sum_{k=1}^{i} n_k/n$ die *Summenhäufigkeitsverteilung* von X.

Für die graphische Darstellung der Häufigkeitsverteilung wird das *Histogramm* (s. Abb.) gewählt. für die Summenhäufigkeitsverteilung die *Treppenfunktion.*

1.1.2 Wichtige Verteilungen

Die Normalverteilung

Die Dichte der Normalverteilung ist gegeben durch

$$g(x) = \frac{1}{2\pi\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.1}$$

wobei μ (Lage) der Mittelwert und σ (Breite) die Standardabweichung der Normalverteilung ist. Durch die z-Transformation lässt sich die Normalverteilung auf die Standardnormalverteilung mit $\mu = 0$ und $\sigma = 1$ transformieren.

Die Normalverteilung bildet die Basis fast der gesamten statistischen Theorie. [1]. Auch bei der Analyse der Microarraydaten werden wir sehr oft von der Annahme der Normalverteilung Gebrauch machen. Allerdings sollten wir uns klarmachen, dass rein experimentell zahlreiche Untersuchungen gezeigt haben, dass die echten Fehler selten, wenn überhaupt normal verteilt sind.

1.2 Schätzung von Parametern

Allgemein erhofft man sich beim Ziehen einer Stichprobe, einen unbekannten Parameter γ der Grundgesamtheit, z.B. den Mittelwert, aus der Stichprobe zu schätzen.

1.2.1 Eigenschaften von Punktschätzungen

[1] "Everyone believes in the normal law, the experimenters because they imagine it is a mathematical theorem, and the mathematicians because they think it is an experimental fact." (Gabriel Lippman, in Poincaré's Calcul de probabilités, 1896)

Chapter 2

Theoretical Background

Here we write about the theoretical background.

2.1 Ordinal Data

2.2 Tangles

Tangles have been a tool in mathematical graph theory, introduced originally by [RS91] with a diverse range of application

In recent times, through the work of [**?**], they have been successfully applied to solve problems of clustering. The mentioned work delivers an algorithmic framework and theoretical guarantees for basic problem settings. Additionally, it delivers simplified notations, adapted to the domain of computer science. When talking about Tangles, we will exclusively use the definitions introduced there, not those that might be common in mathematics.

In this section, we will now deliver a very brief recap of the basic notions, theory and applications of Tangles in a clustering context. For more in-depth explanations of the algorithms and exact procedures, refer to [**?**].

2.2.1 What is a Tangle?

The central object in Tangles Theory is a **bipartition** (which we also refer to as a cut). A bipartition is simply a way of dividing a set of elements $V = \{v_1, v_2, \ldots\}$ into two distinct subsets $A, B \subset V$, such that $A \cap B = \emptyset$ and $A \cup B = V$. We can also write a bipartition as $P = \{A, \overline{A}\}$, with $A \subset V$ and $\overline{A}$ being the complement of A with respect to V.

For such a bipartition to be useful in clustering, we expect it to hold some degree of information about the cluster structure of our data. This means that a good bipartition should not separate groups of data that are tightly coupled. If we imagine a graph data structure, a good bipartition $P = \{A, \overline{A}\}$ might be a separation of the set of nodes V such that there are only a few edges between A and $\overline{A}$. How useful a cut might be for our clustering will be quantified through a **cost function** $c : \mathcal{P}(V) \to \mathbb{R}$, with $\mathcal{P}(V)$ denoting the power set of V. One is free to choose this cost function and it might be dependent on the problem at hand.
Assume that for a set of elements V that we are equipped with a set of bipartitions $\mathcal{B} = \{\{A_1, \overline{A_1}\}, \ldots, \{A_n, \overline{A_n}\}\}$ on V. Coupled with the cost function, this set of bipartitions should tell us a lot about the cluster structure of the data: we know for all bipartitions, how much they do or don't separate dense regions in V. The task of the Tangles framework is to aggregate the information present in the bipartitions and bring it into a useable form. For this, we process $\mathcal{B}$ to a set of so-called **Tangles**, which correspond to specific ways of orienting the cuts in a consistent way such that they point to cohesive structures in the data. Orienting here means that we pick one specific side of a bipartition. An **Orientation** of $\mathcal{B}$ is then a set $O = \{o_1, o_2, \ldots o_n\}$, where o_i corresponds to either the partition A_i (oriented *left*) or $\overline{A_i}$ (oriented *right*). A consistent orientation (which we also call a Tangle) is an orientation for which:

$$\forall A, B, C \in O : |A \cap B \cap B| \geq a. \tag{2.1}$$

for some fixed parameter $a \in \mathbb{N}$, which we refer to as **agreement** parameter. This point is also where we need the cost function: Without it, a lot of reasonably sized sets of bipartitions wouldn't allow for any tangles, as there are simply too many of them to consistently orient. Imagine if our set of bipartitions would contain a few random bipartitions: on average, each of these cuts our set of points in half, so we can at most consistently orient on the order of $O(\log(n))$ many of them. Using the cost function, we can simply restrict our tangles to a set of low-cost (and thus very insightful) cuts P_ψ, using a threshhold $\psi \in \mathbb{R}$ such that $P_\psi = \{Pc(P) \leq \psi\}$. A tangle on P_ψ is then said to have order ψ.

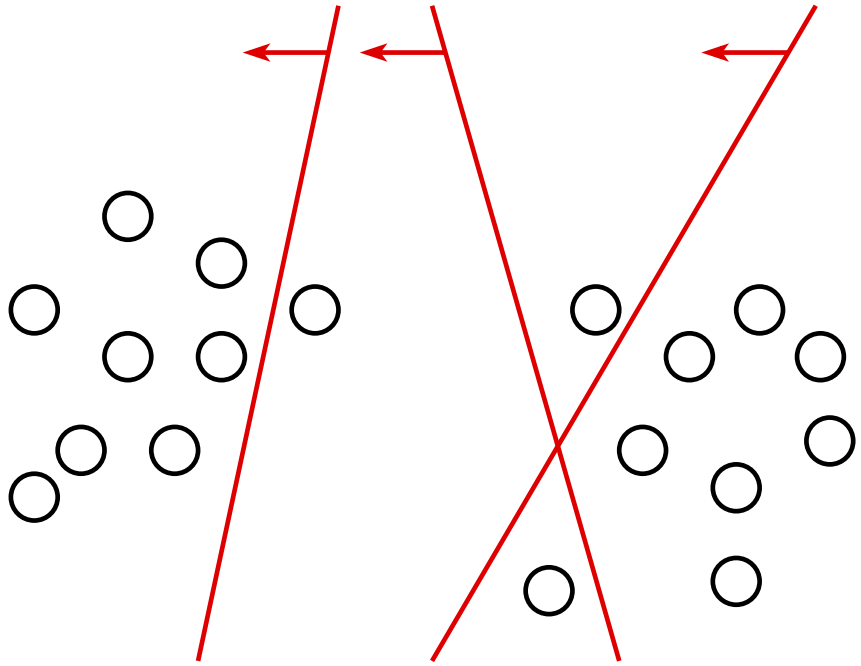

Figure 2.1: An example of how a simple tangle might look like, if we assume a reasonably sized agreement (say $a = 3$). The red lines represent simple cuts, which divide the sets of points into a bipartition. A tangle on this set of bipartitions might orient all bipartitions to the left side (indicated by the arrow), so that they point to the dense structure there. Another possible tangle might orient all cuts to the left. Notice that a tangle on this set of bipartitions can only either point all bipartitions to the left or to the right, else the consistency criterion is violated. This might already give a good intuition on why tangles are able to find dense structures in data.

2.2.2 Processing Tangles to a clustering

As we have seen in Figure 2.1, a tangle might correspond directly to a cluster. But, a given set of bipartitions usually allows for a wide variety of possible tangles, some of them pointing to different or overlapping clusters. We now have to process this set of tangles into a useable clustering. This step is a bit involved and we aim to only give a rough, intuitional overview here.

Given a set of bipartitions $\mathcal{B} = \{b_1, b_2, \dots b_n\}$, we first want to determine for all possible orders ψ the sets of all tangles of order ψ on $\mathcal{B}$ according to a given cost function c. Inuitively, the order of the set of tangles determines how coarse the clustering is they define: If we only use bipartitions with a low cost (so in the case of small ψ), then the bipartitions cut only very loosely connected structures. If the cost is higher, the bipartitions are allowed to cut through more dense regions. This directly induces a sort of hierarchy, where

we go from coarse cluster structures to fine ones with increasing ψ.
To better handle this procedure computationally, we build a tree structure on the set of tangles, called the **Tangle Search Tree**. In the tangles search tree, one node represents a possible tangle. Every level of the tree contains all possible tangles of a certain threshhold ψ_i which directly corresponds to the cost of bipartition b_i. The exact makeup of the tangle is determined by walking the path from the root to the node, and adding bipartition b_i to the tangle in a left-oriented way, if it is a left child and in a right-oriented way if it is a right child. An exemplary tangle search tree is illustrated in **??**.

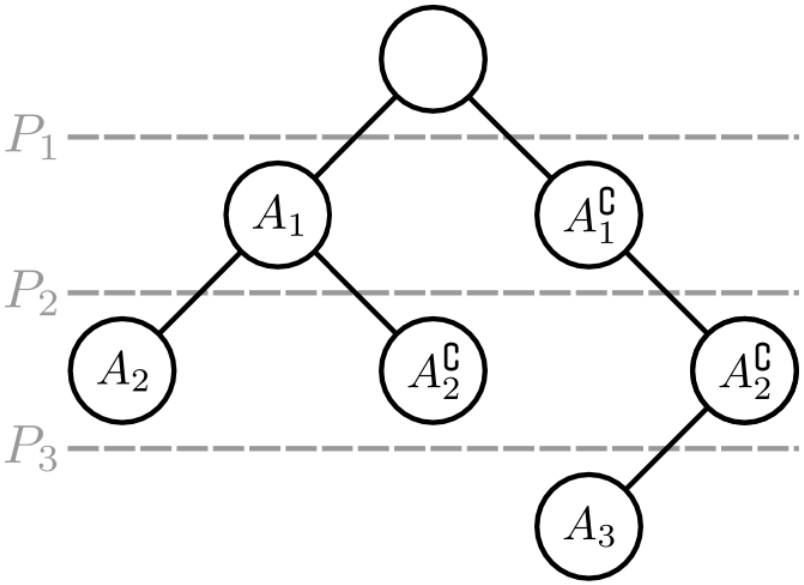

Figure 2.2: An example of a possible tangles search tree for a set of bipartitions $\mathcal{B} = \{\{A_1, \overline{A_1}\}, \{A_2, \overline{A_2}\}, \{A_3, \overline{A_3}\}\}$. Each level corresponds to the tangles of the order given by the bipartition P_i that is indicated to the left of it. Let us take a look at the sole node in level 3. If we would want to find out, which orientations the corresponding tangle T consists of, we just walk from the root to it and add the bipartitions in the direction indicated by the tree. We end up with $T = \{\overline{A_1}, \overline{A_2}, A_3\}$. Figure taken with permission from **??**.

We can now obtain a soft, hierarchical clustering from this tangle search tree. For this, the interesting nodes are those where the tree splits up (as this represents a new clustering) and the leaves (which correspond to the clusters). For each of the *splitting nodes*, we determine the set of *characterizing cuts*. A cut belongs to this set, if it is both oriented the same way inside the subtrees, and if it is oriented in a different way between the two subtrees. To illustrate this, we take a look at the exemplary tree in **??**. Here, for the root node, P_1 is a characterizing cut (pretty trivially), while P_2 is not: below the node A_1, the bipartition is both oriented to the left and to to the right, violating the

requirement that the cuts are always oriented the same way inside the subtrees.

The characterizing cuts express some kind of agreement on how to align the bipartitions in the subtrees of the node, which we can leverage for a soft clustering. If we now want to determine, with what probability a point a belongs to a cluster C represented by a leaf in the tree, we simply walk down the tree from the root, and count at every splitting node how many characterizing cuts contain the point a, and how many don't. Normalized by the total amount of characterizing cuts, we can interpret this as a probability to walk down either the left or the right path. To get a total probability that a belongs to C, we simply multiply the probabilities that we obtain on the path to C at every splitting node.

2.3 Clustering Ordinal Data

Chapter 3

Simulations

Here we write about the simulations we did.

Chapter 4

Real-world data

here we write the conclusion

Appendix A

Further Tables and Figures

Viele Arbeiten haben einen Appendix. Besondere Sorgfalt muss beim Nummerieren der Tabellen und Abbildungen gewährleistet sein.

Nummer	Datum
1	1.1.80
2	1.1.90

Table A.1: Erste Appendix-Tabelle

Nummer	Datum
1	1.1.80
2	1.1.90

Table A.2: Zweite Appendix-Tabelle

Bibliography

[RS91] Neil Robertson and P. D Seymour. Graph minors. X. Obstructions to tree-decomposition. *Journal of Combinatorial Theory, Series B*, 52(2):153–190, July 1991. The paper that is cited in the tangles paper as basis for tangles.

Selbständigkeitserklärung

Hiermit versichere ich, dass ich die vorliegende Masterarbeit selbständig und nur mit den angegebenen Hilfsmitteln angefertigt habe und dass alle Stellen, die dem Wortlaut oder dem Sinne nach anderen Werken entnommen sind, durch Angaben von Quellen als Entlehnung kenntlich gemacht worden sind. Diese Masterarbeit wurde in gleicher oder ähnlicher Form in keinem anderen Studiengang als Prüfungsleistung vorgelegt.

Ort, Datum Unterschrift