修 士 論 文 の 和 文 要 旨

研究科・専攻	大学院　情報理工学研究科　情報・ネットワーク工学専攻　博士前期課程		
氏　　名	原田　貴史	学籍番号	1631121
論 文 題 目	パーシステントホモロジーによる多次元ファジィ集合の同定		

要　旨

ディープラーニングなどを用いて多変量かつ多次元なデータを解析したデータはブラックボックスになっている問題がある。これは人が理解しデータを編集する上で大きな阻害になっている。そこで、多変量かつ多次元データを人の感覚を反映したデータへと解析する手法として、多次元ファジィ集合がある。

既存手法では、データのネットワーク構造を作成することで、幾何学構造を保存し、ネットワーク構造上でのデータの分布密度を計算して多次元ファジィ集合を生成している。人の感覚を反映するために、サンプル密度を計算するための閾値ϵ_Mと、幾何学構造を保存するための閾値ϵ_Fを決定する必要がある。既存手法では手動で模索し閾値を導出していた。

本研究では、パーシステントホモロジーを用いて、幾何学構造を同定し、閾値ϵ_M, ϵ_Fを導出することで自動で多次元ファジィ集合を生成し、手動で求めた閾値で生成した多次元ファジィ集合と比較検討を行った。

ホモロジー群は、単体複体と呼ばれるものに対して、各次元での大域的なつながり具合を表すものであり、1 次元の連結成分、2 次元の穴、3 次元のわっか、などの「穴」に関連した幾何学的意味を持つ。パーシステントホモロジーは、単体複体の増大列を用いることで、ホモロジー群xが生成された時間bと消滅した時間dを計算する手法である。これにより、対象のホモロジー群xの生成元の遷移を特徴づけることができる。

本研究では、パーシステントホモロジー群の中の生存時間が短い要素をノイズと考え、重み関数$w(x)$による重み付けを行った。二つの閾値ϵ_M, ϵ_Fを、重みの付いたパーシステントホモロジー群から 0 次元ホモロジーを除き、生存時間の中点の重み付き平均$\epsilon_M = \sum_{i=1}^{N} w(x_i)(b_i + d_i)/2\sum_{i=1}^{N} w(x_i)$と消滅時間の重み付き平均$\epsilon_F = \sum_{i=1}^{N} w(x_i)d_i/\sum_{i=1}^{N} w(x_i)$とした。これにより、データの幾何学構造を保存しつつ、閾値を導出することができた。導出した閾値と手動で求めた閾値を用いて比較実験を行い有用性を確認した。

電気通信大学情報理工学研究科
情報・ネットワーク工学専攻情報数理工学プログラム修士論文

パーシステントホモロジーによる
多次元ファジィ集合の同定

平成 30 年 3 月 13 日

情報数理工学プログラム

学籍番号 1631121

原田貴史

指導教員 緒方 秀教
村松 正和
助言 西野 順二

目 次

第1章　序論

1.1　研究の背景と目的

近年、計算機の進歩によりビッグデータという複雑で膨大なデータが得られるようになってきた。このビッグデータを活用するために、ディープラーニングといった機械学習による解析が行われているが、解析手法によっては解析したデータが、人が見ても理解できないブラックボックスになっている場合がある。ビッグデータの解析を理解し有効活用するには、人の感覚で理解できるような手法が必要である。

人の感覚をシステムに反映する理論として、ファジィ理論 [1][2] というものがある。ファジィ理論は人が関係するシステムを運用するために、ザデーが 1965 年に提案したものである。ファジィ理論の基礎となるファジィ集合では、データがある集合に属する度合いであるメンバーシップ値というものが与えられる。このメンバーシップ値によってデータのあいまいさやデータに対する人の感覚が示される。ファジィ集合は多くの場合、一次元のデータから変換されて生成されており、従来の多次元データのファジィ集合は、一次元ファジィ集合とその直積によって表されてきた。しかしながら、この方法では多次元かつ多変量空間においては、組み合わせが多くなり複雑なものを表すのが困難になっている。これに対し糟谷らが多次元な問題に対して自然な多次元のファジィ集合を定義する方法 [3][4][5] を提案している。問題の表現空間の変数が多数であるような複雑問題に対する、ファジィ集合によるアプローチをこのへんファジィと呼ぶ [6]。

糟谷らが提案した手法は、モデリング対象のランダムサンプリングデータをもとに幾何学的な構造を保存しながら空間中のサンプル密度にもとづきファジィ集合を構成する手法である。この手法では、サンプル点同士の近傍点をパスでつなぐことで、ネットワーク構造を構成し、幾何学的構造を保存している。この手法により、n 次元ユークリッド空間に分布したサンプル点に大域的な幾何的構造を与えることができ、幾何学的構造を保存することで、より人の感覚に近い多次元ファジィ集合ができると考えられている。

しかし、糟谷らが提案した手法では、幾何学構造を保存するための閾値 ϵ_M を手動で求める必要がある。手動で編集することで、人の感覚に近い幾何学構造を保存することができるが、人が多次元空間を認識することは難しく、その調整が困難である。また、閾値の値によっては全点が孤立してしまう場合や、全点同士が結合してしまうようなネットワークを作ってしまい、ネットワーク構造を構成する意味がなくなってしまう可能性がある。

また、この手法では、サンプル密度を計算するための閾値も手動で編集する値となっている。サンプル密度は、ある点における自身を含めた近傍点の個数を、全点における近傍点の個数の最大値で除したものであり、これをファジィ集合に属する度合いであるメンバーシップ値としている。よって、このサンプル密度を計算する閾値によってファジィ集合の値は決定づけられる。幾何学構造を保存するための閾値と同じく、手動で編集することによって人の関係する集

合として、人の感覚を反映しやすくなるが、この閾値も問題に対して、極端に大きくもしくは小さい値をとった場合に、サンプル密度が全点において同じもしくは一部の点に大きく偏ることがある。これはデータに対して人の感覚を反映できていない可能性があり、またシステムとして扱い辛い集合となる。

これに対し、原田らによってパーシステントホモロジーによる多次元ファジィ集合の生成が提案されてる [7][8]。この手法では最もよく現れる幾何学構造しか保存できておらず、データに含まれるより多くの幾何学構造を考慮する必要があると考えられる。

ここまでの経緯をまとめると以下のようになる。

1. 多次元ファジィ集合の提案(西野)
2. サンプル密度による自動生成
3. ネットワーク距離による改善(糟谷)
4. パーシステントホモロジーによる改善(本研究　原田)

よって、本研究の目的は、先行研究の問題点を改善するために、二つの閾値、幾何学構造を保存するための閾値、サンプル密度を計算するための閾値に対して、妥当な値を導出する方法を提案することである。

1.2　研究の概要

糟谷らが提案した手法における、手動で決定する二つの閾値の問題点はそれぞれ、値の取り方によって極端なデータ構造を生成してしまうことである。そこでデータの幾何学的構造を同定し、そこから二つの閾値を導出する方法を提案する。

幾何学構造を同定する手段として、ホモロジー群がある。ホモロジー群はデータの各次元に対応して、1 次元では連結数、2 次元では穴、3 次元ではわっかなど、各次元の穴に相当する幾何学的特徴を捉えるものである [9][10]。本研究で用いるサンプル点データの場合、ホモロジー群を計算する際に、n 次元空間の一般の位置にある $n+1$ 個の点の集まりによる単体を生成し、その単体の集まりである単体複体を構成する。計算ホモロジー群はこの単体複体と穴を計算手続きによって導出する。本研究の対象データはサンプル点データである、よって単体を生成する際、ある点とどの位置にある点の集まりを単体としてみなすか閾値を決める必要がある。そこで、その閾値の変化とホモロジー群の遷移を確認する計算手法としてパーシステントホモロジー [11][12] を用いる。パーシステントホモロジーでは、どの異なる点同士を単体するかという閾値を、ある初期値からあるタイムステップ毎に増加させていき、それぞれの閾値でのデータのホモロジー群を計算する。この手法によって、ホモロジー群の値がどの閾値で生成され初め、どの閾値では消滅したのか確認することができる。これによって、単体を生成する閾値の変化と幾何学構造との関係がわかる。このようなホモロジー群の生成と消滅を表したものをパーシステント図と呼ぶ。

パーシステント図では、一部のホモロジー群では生成されたタイムステップと、消滅したタイムステップが非常に近いものも生成される。この点はノイズである可能性が高いため、他の

点と区別する必要がある。よってこのノイズを区別するために先行研究の手法を用いて重み付けを行う。そして、重み付けを行われたパーシステント図に対して、ホモロジー群の生成と消滅時刻の重み付き平均を計算し、その点を幾何学構造を保存するための閾値とする。同様に、ホモロジー群の消滅時刻の重み付き平均を計算し、その点をサンプル密度を計算する閾値とすることを提案した。

第2章　多次元ファジィ集合

2.1　ファジィ理論

ファジィ理論とは、ファジィ集合、ファジィ論理、ファジィ測度をコアとする理論枠組みのことである。

人間の思考過程の重要な要素は言語であるといってよい。言語なくしては人間は考えることができず、思考の系列というものがあるとすれば、それは言語の系列である。この言語は数学記号などと異なって、あいまいさをもつものである。このあいまいさがファジィ理論でファジィネスと呼ばれているものである。

ファジィネスとはたとえば、「若者」や「大きい」に見られる言葉の意味とか概念の定義のあいまいさのことである。これまでのシステムの不確かさは、確率的不確かさであり、「明日、雨が降る」ことの偶然性などである。

ファジィ理論においてはシステムの記述言語としてファジィネスを含むものを使い、このことによってシステムの複雑性とデータの不完全性に対処する人間の認識、判断、思考などの術をあらわそうとするのである。

たとえば、右に曲がるカーブを車が抜けるとき、「カーブに近づいたら徐々に減速し、ハンドルをゆっくり右に切る」というような制御アルゴリズムをファジィ制御では用いる。

2.2　ファジィ集合

ファジィ集合とは、言葉の意味や概念の定義にみられるあいまいさを定量的に表すための集合概念である。

熱い一冷たい、近い一遠い、などの言葉の意味は事物の置かれた状況と、判断する人間の感覚に依存して、とてもあいまいである。われわれが日常的に使っている自然言語は数値を用いない定性的なものであるが、これらの例にみられる言葉はなにかしら量に関係した、比較的に定量化しやすいものであり、ファジィ集合はさしあたってこの種の言葉のあいまいさを扱うものと考えてよい。

言葉のもつ意味のあいまいさの定量化を「風呂の湯が熱い」というときの「熱い」を例にとって考えてみる。図 2.1 は「熱い」という言葉の意味を関数を用いて定量化したものである。横軸には風呂の湯の温度を 32℃から 48℃までとってある。縦軸は 0 から 1 までのグレードと呼ばれる数値を示している。横軸は言葉の意味を考えるときの量の世界であり「台」と呼ぶ。「熱い」といっても風呂とお茶では変わってくるので台の範囲は対象としている世界によって変化し、また縦軸のグレードも対象とする人によって変化する。このグラフのグレードは湯のそれぞれの温度に対して、その温度がどの程度「熱い」とみなせるかという度合いを示している。

図 2.1 ではグレード の値は湯の温度が 38 ℃なら 0.3、44 ℃なら 0.9 をとっており 人間の「 熱い」に対するあいまいな思考を 表現できている。グレード の最大値が 1、最小値が 0 という 数値そのものはとくに意味をもつわけではない。

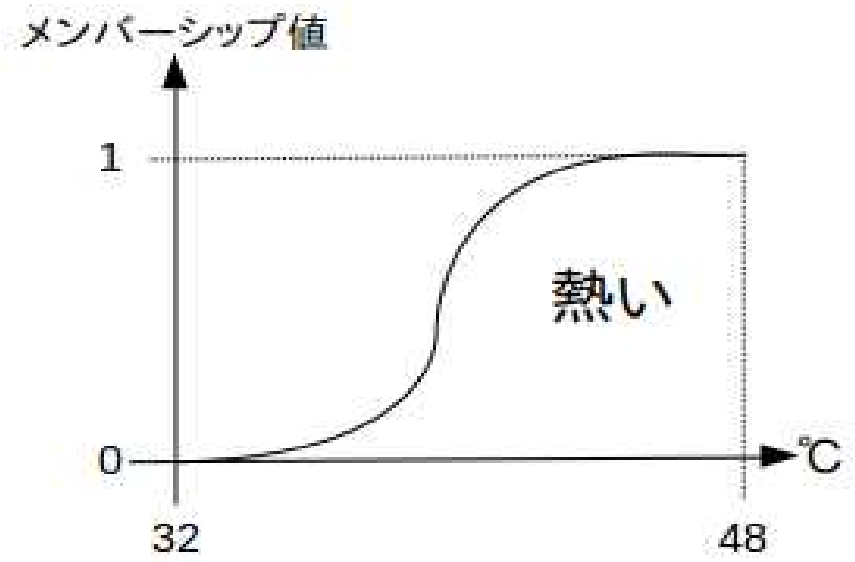

図 2.1: ファジィ 集合

このように、意味の定量化は台の範囲と 台の上のグラフの形の 2 つに依存する。台の範囲は客観的、グラフは主観的という 比較ができる。例からわかるように、言葉の意味のあいまいさは、グレード が 0 か 1 の間のいろいろな 値をとるというところに現れている。

2.3 多次元ファジィ 集合

多次元ファジィ 集合は、 $\mu_A(x), x \in \mathbb{R}^n$ のメンバーシップ関数で特徴付けられた任意の n 次元パラメータ空間 $\mathbb{R}^n$ 上の曖昧な部分集合として定義される。n 次元上のファジィ 集合は、これまでは一次元ファジィ 集合の直積又は和によって表されてきた。この方法では多次元空間上において多数のファジィ 集合が必要となり 集合が複雑になってしまい、シンプルで精度を 良く表現することは難しくなる。そこで多次元ファジィ 集合を 直接定義することで多次元空間上でも 自然に 表現することができる。

2.4 既存手法

先行研究における、多次元ファジィ 集合の生成方法を 示す。この手法では 2 つの閾値を 用いて、多次元ファジィ 集合を 生成する。

2.4.1　幾何学的構造を保存するための閾値 ϵ_M

1. パラメータ空間 $\mathbb{R}^n$ のサンプル点の集合を $D = \{x_1, x_2, \ldots, x_N\}$ とする。

2. 近傍閾値 ϵ_M は、ある 2 点 x_i, x_j のユークリッド距離 $d_{i,j}$ とし、定数 $K_m \in \mathbb{R}$ を用いて、式 (2.1) のように定義する。

$$\epsilon_M = \frac{\sum_{i=1}^{N} \min_{j=1}^{N}(d_{i,j})}{N} \times K_m. \tag{2.1}$$

3. 二点 $x_i, x_j \in D$ のユークリッド距離が、近傍閾値 ϵ_M 以下の組を連結した近傍グラフを構成する。

4. 任意の二点 x_i, x_j のネットワーク距離 d_{p} を、近傍グラフ上の最短パスによって定義する。

2.4.2　サンプル密度を計算するための閾値 ϵ_F

1. 近傍閾値 ϵ_F は、ある 2 点 x_i, x_j のネットワーク距離 $d_{\mathrm{p}i,j}$ とし、定数 $K_f \in \mathbb{R}$ を用いて、式 (2.2) のように定義する。

$$\epsilon_F = \frac{\sum_{i=1}^{N} \min_{j=1}^{N}(d_{\mathrm{p}i,j})}{N} \times K_f. \tag{2.2}$$

2. ある点 x におけるサンプル密度は、ネットワーク距離閾値 ϵ_F 以下の距離にある点の個数 $N_{\mathrm{n}}(x)$ を、$N_{\mathrm{n}}(x)$ の最大値 $N_{\max}$ で除して、これを当該点のメンバーシップ値 $\mu(x)$ とする。

$$\mu(x) = \frac{N_{\mathrm{n}}(x)}{N_{\max}}. \tag{2.3}$$

本研究では、先行研究における式 (2) の ϵ_M,ϵ_F を導出する計算について、パーシステントホモロジー群を使用しモデルを同定することを提案する。

2.5　各閾値の算出

糟谷らが提案した多次元ファジィ集合生成方法の課題は、二つの閾値 ϵ_M,ϵ_F の設定方法である。二つの閾値 ϵ_M,ϵ_F は人の感覚を多次元データに表すための重要なパラメータである。

2.5.1　課題 1:幾何学構造を保存するための閾値 ϵ_M

幾何学構造を保存するための閾値 ϵ_M は、幾何学構造を保存することでより人間の感覚に近い多次元ファジィ集合を生成することができるという考えから設定されている。幾何学構造を保存しなければいけないが、その値を手動で調整する必要がある。

例として、図 2.2 の 2 次元空間上のサンプリング点データを用いた場合、人が修正したものを図 2.3 に示す。図 2.3 において、z 軸はメンバーシップ値を表している。

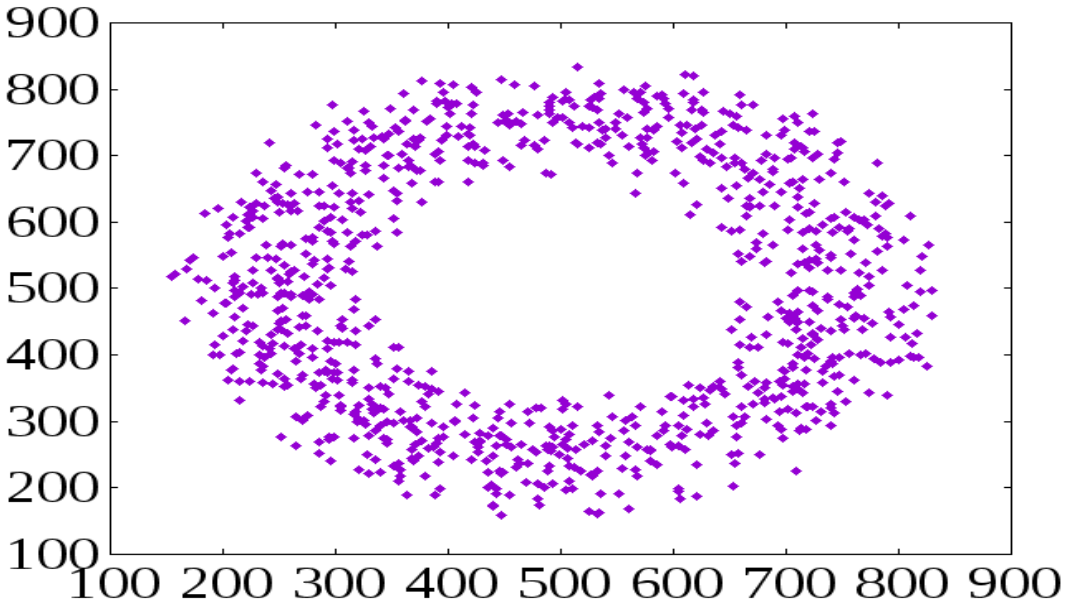

図 2.2: サンプル点データ

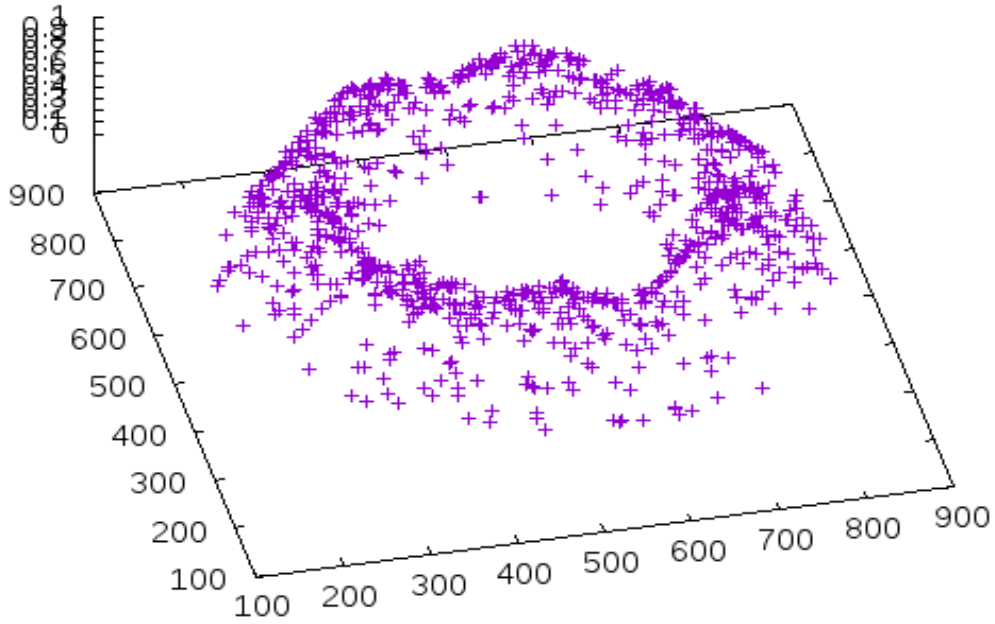

図 2.3: サンプル点データの多次元ファジィ集合

このサンプリング点データでは円形の幾何学構造があることがわかることから、ドーナツ上にサンプル密度が取れていれば良いことがわかる。よって、多次元ファジィ集合を生成した際に円の内側のサンプル密度が高ければ、ネットワーク構造が円の穴を保存していないことがわかり、サンプル密度のばらつきが大きければ個々の点が孤立してしまいネットワーク構造を保存できていないことがわかる。このように、2次元までのデータなら多次元ファジィ集合にした際に確認し、修正することも容易である。また、ネットワーク構造を構成した際の無向グラフを作成することで、3次元までのデータの幾何学構造を生成できているか確認することができる。図2.4,2.5,2.6に閾値 $K_m = 3, 9, 30$ と調整した場合の無向グラフを示す。

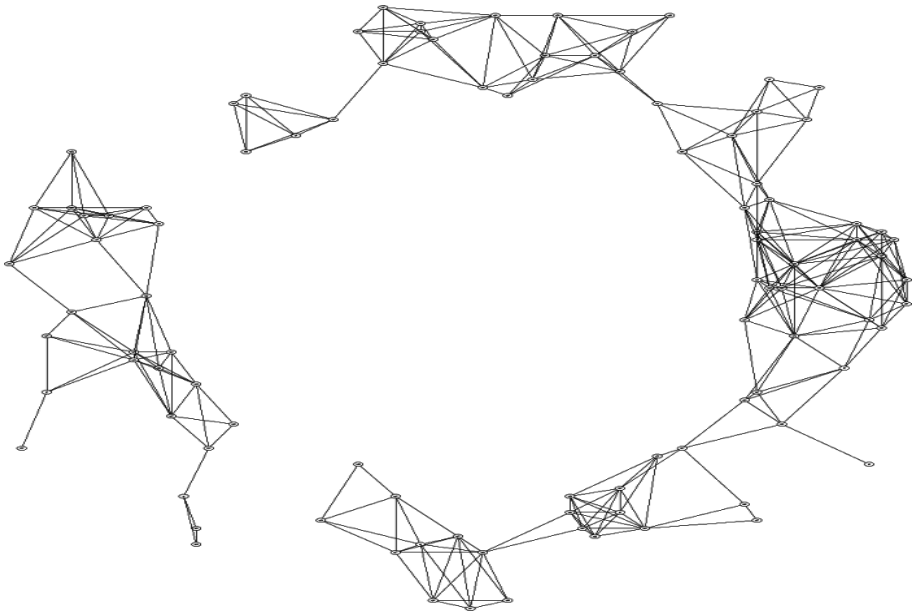

図 2.4: $K_m = 3$ の無向グラフ

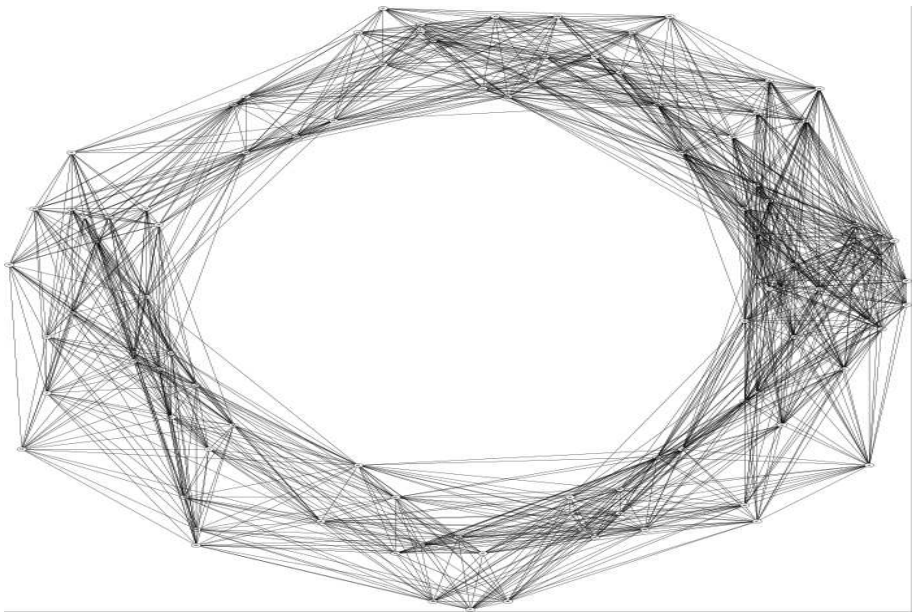

図 2.5: $K_m = 9$ の無向グラフ

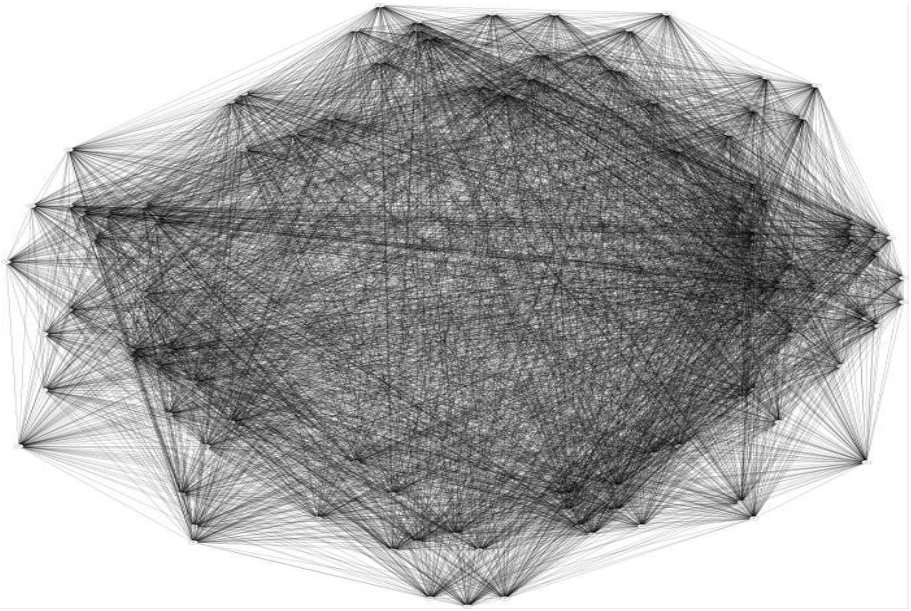

図 2.6: $K_m = 30$ の無向グラフ

このように無向グラフを確認することで手動で調整することができる。この場合は、$K_m = 9$ の無向グラフが最も円の幾何学構造を保存できていると考えられる。このようなサンプル点が持つ何かしらの幾何学構造を計算によって保存できることが人の感覚に近い多次元ファジィ集合を作成しやすくすると考えられる。また、人の認識できない多次元空間上でも同様に計算によって幾何学構造を同定し多次元ファジィ集合を作成することで、人の感覚に近い多次元ファジィ集合が作成できると考えられる。

2.5.2 メンバーシップ値の分布分析

メンバーシップ値の分布を確認することで、多次元ファジィ集合を編集する方法がある。

例として、極端な構造をとらないために、メンバーシップ値の分布を参考にする方法をあげる。ここでは、ロボカップシミュレーション 2D のボールの位置のログデータを用いた。$K_f = 100$ に固定し、$K_m = 1, 5, 10$ と変化させた場合の多次元ファジィ集合のメンバーシップ値の分布を図 2.7~2.9 に示す。

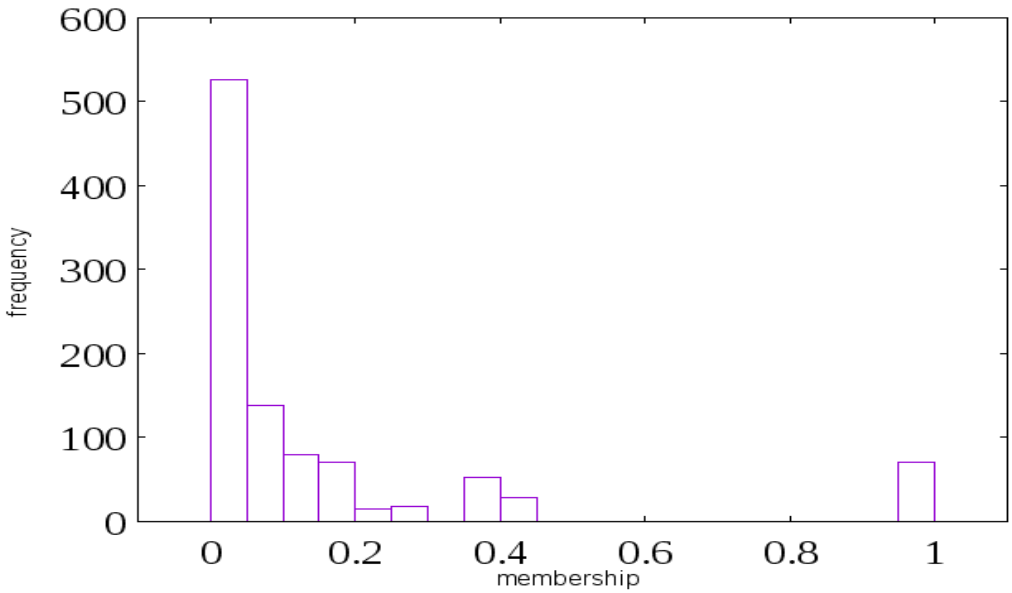

図 2.7: $K_m = 1$ の多次元ファジィ集合のメンバーシップ値の分布

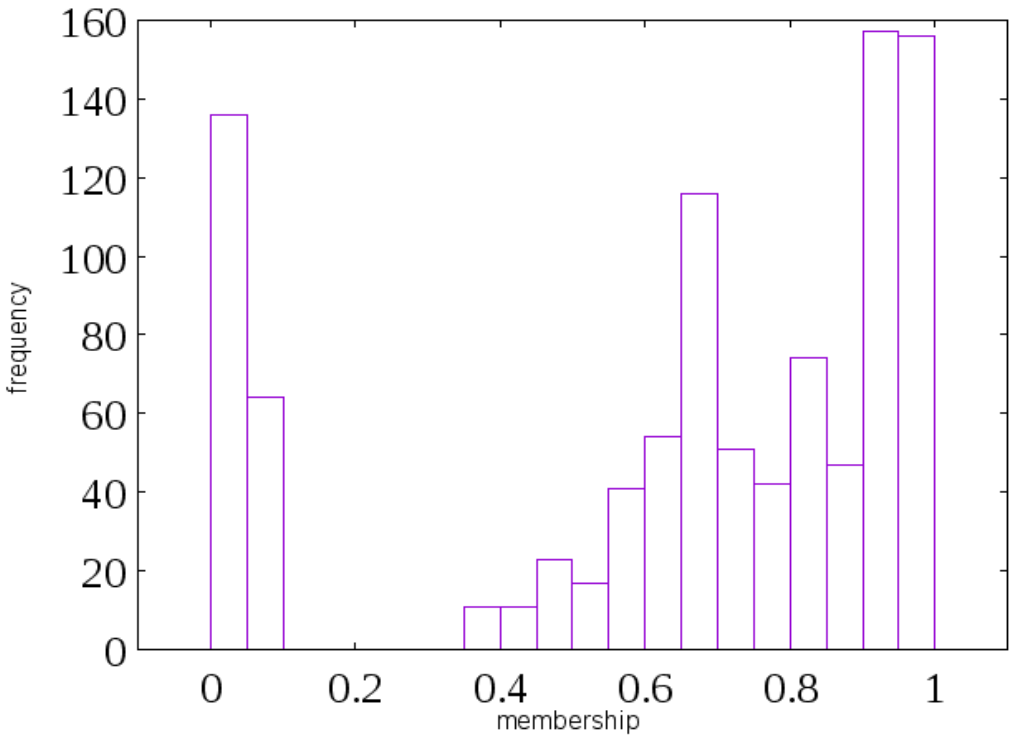

図 2.8: $K_m = 5$ の多次元ファジィ集合のメンバーシップ値の分布

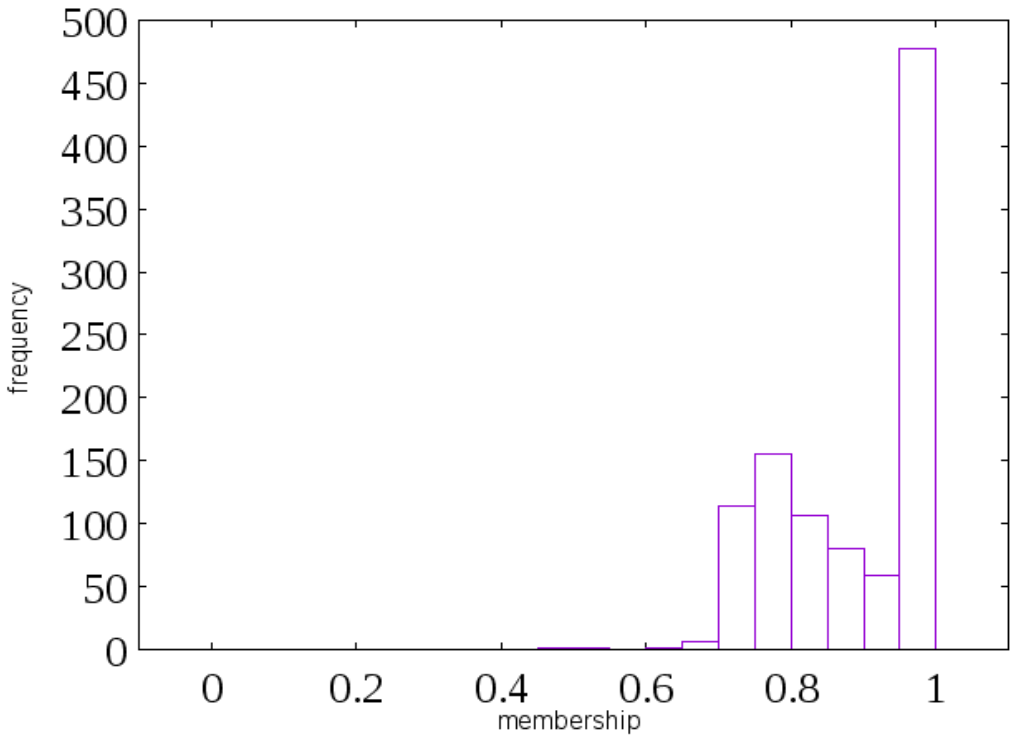

図 2.9: $K_m = 10$ の多次元ファジィ集合のメンバーシップ値の分布

このようにメンバーシップ値の分布の偏りよりから、多次元ファジィ集合の極端なデータ構造を把握することができる。しかしこの手法は、元々大きく偏っているようなデータの構造は把握することができないため、一意にこの分布から予測できるわけではない。

この分布は、システムで使用する観点から見ると、データがバラついていた方がデータを比較しやすいが、一様に分布していた場合、分布が偏っていた方が正しいとも考えられ、多方面から見る必要がでてくる。ネットワーク構造のエッジの数から検討する手法を用いた場合でも、偏りなどから幾何学構造は検討することはできないと考えられる。

よって、多次元データに対して、手動で調整することも難しくために、幾何学構造を保存するための指標がなければ、構造を保存できずに多次元ファジィ集合を構成してしまい、結果として人の感覚を表現できていないものとなると考えられる。

2.5.3 課題 2:サンプル密度を計算するための ϵ_F

サンプル密度を計算するための閾値 ϵ_F は、サンプル点データを見た人が密度の高い点をその集合に属する度合いが高くなると考え、密度が低い点は属する度合いが低くなるという考えられている。よって的確なサンプル密度を取ることが人の感覚を反映した多次元ファジィ集合を生成すると考えられる。

極端に大きな閾値をとってしまうと、サンプル密度は均一になり、小さな値をとると、個々のデータのばらつきが大きくなり、人の感覚をデータに反映できたとはいいづらくなる。しかしながら、人の感覚とは一般的に決められるものでもなく、個々の考えが反映されるため正解

もない。図 2.10 のサンプル点データに対して、閾値 $K_m = 3$ に固定し、閾値 $K_f = 3, 30, 300$ と変化させた場合の多次元ファジィ集合を図 2.11~2.13 に示す。

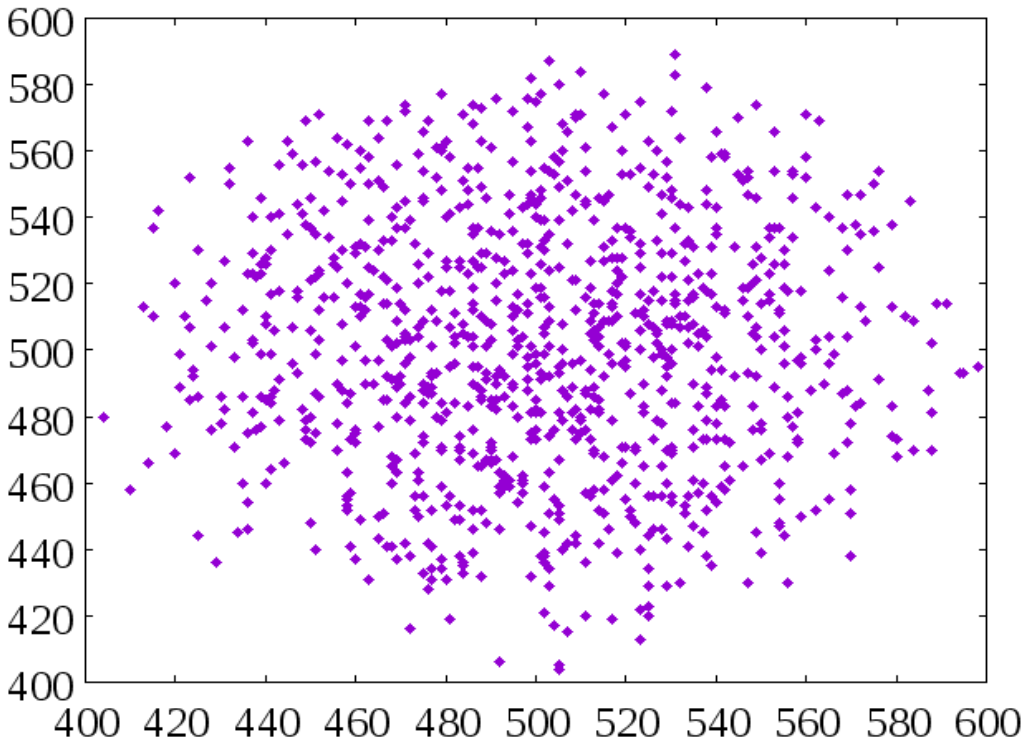

図 2.10: サンプル点データ

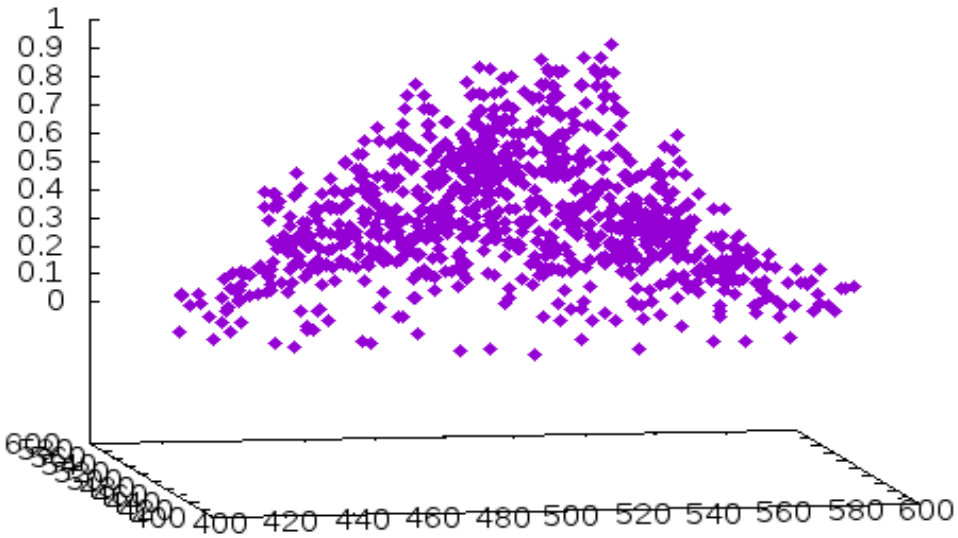

図 2.11: 閾値 $K_f = 3$ の多次元ファジィ集合

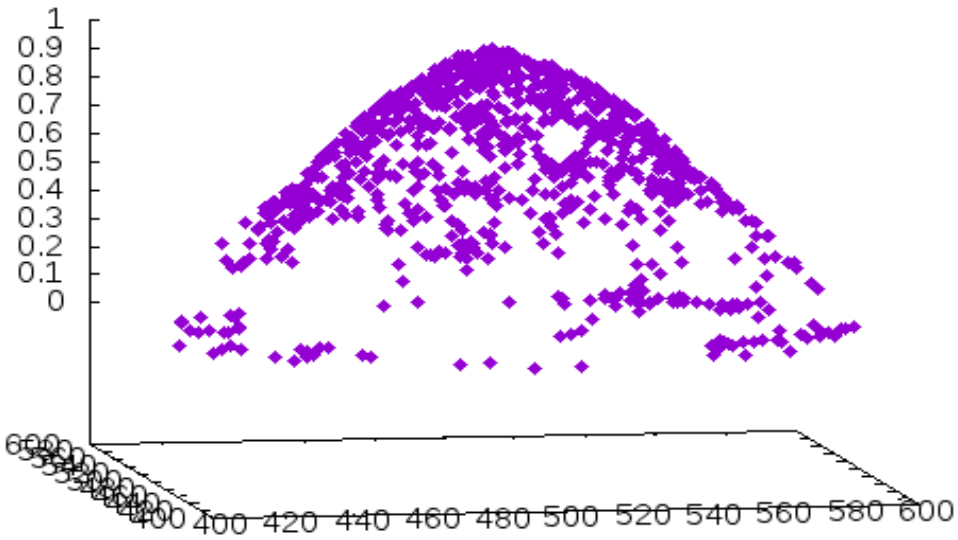

図 2.12: 閾値 $K_f = 30$ の多次元ファジィ集合

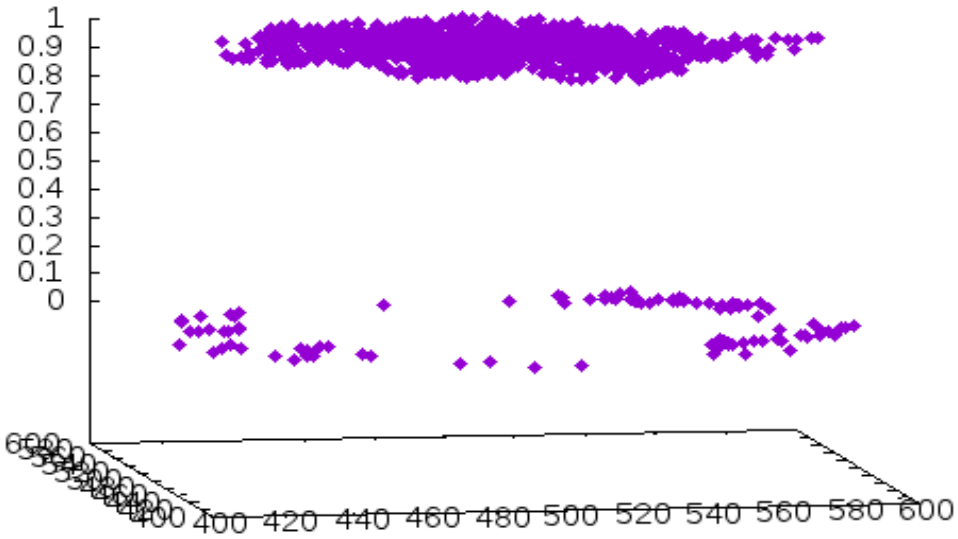

図 2.13: 閾値 $K_f = 300$ の多次元ファジィ集合

このように閾値を変化させることでサンプル密度は変化する、値の取り方によっては分布密度は一定になってしまうことがある。この閾値に対して、人の視認できない多次元データで調整することは難しいと考えられる。人の感覚を反映するには、その多次元ファジィ集合がどの

ような構造でどのようなサンプル密度を取っているか確認しなければ、調整することもできない。しかしながら、この閾値においても全ての点が閾値の範囲内になる閾値やすべての点が孤立するような閾値は、どのようなサンプル点データにおいても全点のサンプル密度が同じになり多次元ファジィ集合の意味を成さないと考えられる。よって、そのような閾値をとらない範囲を決定する必要がある。

また、データの種類によって閾値の変化と関係を調査することで、閾値の指標を検討することができる。

第3章　パーシステント ホモロジー

多次元ファジィ集合生成方法の二つ閾値を導出するために、対象の幾何学構造を導出することのできるホモロジー群を計算する。ホモロジー群は単体複体と呼ばれる点の集まりに対して、各次元での大域的なつながり具合を表すものである。

まず、ホモロジー群を計算するため、点からの単体複体する必要があり、点から単体複体を作成するには、点と近傍点を集まりとみなす必要がある。この近傍点の決め方によってホモロジー群は変化してしまう。

そこで、この単体複体の変化に対するホモロジー群の遷移を計算できるようにするのが、パーシステント ホモロジー群である。単体複体のフィルトレーション (増大列) に対して、パーシステント ホモロジー群を適用することで、各閾値に対する各次元のホモロジー群の生成と消滅を確認できる。

パーシステント ホモロジー群とは、対象のホモロジー群の生成元の遷移を特徴づける方法である。

3.1　ホモロジー群

n 次元パラメータ空間 $\mathbb{R}^n$ 内の $k+1$ 個の点 $x_0, x_1, \ldots, x_k$ が k 個の一次独立なベクトル $\overrightarrow{x_0x_1}, \ldots, \overrightarrow{x_0x_k}$ を与える時、それらの点で構成される最小の凸集合を k 単体と呼ぶ。

$\mathbb{R}^n$ 内の有限個の単体の集まり K とする。K が以下の二つの条件を満たす場合、K を単体複体と呼ぶ。

1. K に属する単体 τ の面 $\sigma \prec \tau$ もまた K に含まれる。
2. 2つの単体 $\tau, \sigma \in K$ に対して、$\tau \cap \sigma$ が空集合でないならば、$\tau \cap \sigma$ は τ の面かつ σ の面である。

ホモロジー群は、単体複体に対して、各次元での大域的なつながり具合を表すものであり、1次元の連結成分、2次元の穴、3次元のわっか、などの「穴」に関連した幾何学的意味を持つ。

単体の頂点に順序づけしたものを σ と呼ぶ。K を n 次元単体複体とし、そのすべての k 単体の集まりを

$$K_k = \{\sigma_1, \cdots, \sigma_{n_k} \in K | \dim\sigma_i = k\}. \tag{3.1}$$

で表す。

各 $0 < k < n$ ごとに K_k で生成される自由 Z 可群 $C_k(K)$ は

$$\begin{aligned} C_k(K) &= \mathbb{Z}\langle K_k \rangle \\ &= \{c = \sum_{i=1}^{n_k} \alpha_{\sigma_i} \langle \sigma_i \rangle | \alpha_{\sigma_i} \in \mathbb{Z}\}. \end{aligned} \tag{3.2}$$

で表す。

また、各 C_k ごとに、境界作要素 $\partial_k : C_k(K) \to C_{k-1}(K)$ を向きづけられた単体ごとに

$$\partial_k \langle \sigma \rangle = \sum_{i=0}^{k} (-1)^i \langle \upsilon_0 \cdots \hat{\upsilon_i} \cdots \upsilon_k \rangle. \tag{3.3}$$

で定める。

ここで、一般の k 鎖 $c = \Sigma_{i=1}^{n_k} \alpha_{\sigma_i} \langle \sigma_i \rangle$ については線形拡張

$$\partial_k \langle c \rangle = \sum_{i=1}^{n_k} \alpha_{\sigma_i} \partial_k \langle \sigma_i \rangle. \tag{3.4}$$

で定める。

式 (5),(7) 導入した鎖群と境界作要素からなる系列を、K の鎖複体と呼ぶ。

ここで、k 鎖群 $C_k(K)$ の 2 つ部分加群を導入する。

$$\begin{aligned} Z_k(K) &= \mathrm{Ker}\partial_k \\ &= \{c \in C_k(K) | \partial_k(c) = 0\}, \\ B_k(K) &= \mathrm{Im}\partial_{k+1} \\ &= \{c \in C_k(K) | c = \partial_{k+1}(c'), c' \in C_{k+1}(K)\}. \end{aligned} \tag{3.5}$$

この二つの部分加群を使用すると、単体複体 K の k 次ホモロジー群は、剰余加群

$$H_k(K) = Z_k(K)/B_k(K). \tag{3.6}$$

で定められる。

3.2 パーシステント ホモロジー群

本研究では、単体複体の生成する場合、要素同士を接続するために点間のユークリッド距離を使用する。その要素間の距離の増大させることによる、単体複体の増大列をとったものが、フィルトレーションとなる。パーシステント ホモロジー群とは、単体複体のフィルトレーションから、それぞれのホモロジー群を生成することであり、各次元ホモロジー群の生成消滅を確認することで、対象のホモロジー群の持続性やロバスト性を考察できるようにする方法である。

ここで、単体複体 $K^t, t = 0, 1, \cdots$ のフィルトレーション

$$\mathbb{K} : K^0 \subset K^1 \subset \cdots \subset K^t \subset \cdots . \tag{3.7}$$

を考える。ここでフィルトレーション内の単体複体 K^t を指定する添字 t を時刻と呼ぶ。フィルトレーション $\mathbb{K}$ は、非負整数 Θ が存在し、$K^j = K^\Theta, J \leq \Theta$ が成り立つとき、有限型であるという。またこの性質を満たす Θ の最小値を、フィルトレーションの飽和時刻と呼ぶ。

このフィルトレーションに対して、k 次パーシステントホモロジー群 $PH_k(\mathbb{K})$ は

$$PH_k(\mathbb{K}) = Z_k(\mathbb{K})/B_k(\mathbb{K}). \tag{3.8}$$

で定められる。

また、$Z_k(\mathbb{K}), B_k(\mathbb{K})$ は斉次部分加群なので、パーシステントホモロジー群は次数付き $\mathbb{Z}_2[x]$ 加群として

$$PH_k(\mathbb{K}) = \bigoplus_{t\geq 0} Z_k(K^t)/B_k(K^t) = \bigoplus_{t\geq 0} H_k(K^t). \tag{3.9}$$

で与えられる。

パーシステントホモロジー群 $PH_k(\mathbb{K})$ は、$Z_k(\mathbb{K})$ のある斉次基底 $g_i, \ldots, g_m$ を用いて

$$PH_k(\mathbb{K}) = \bigoplus_{i=0}^{s} \langle [g_i] \rangle \oplus \bigoplus_{i=s+1}^{s+r} \langle [g_i] \rangle, \tag{3.10}$$

$$\mathrm{Ann}([g_i]) = (x^{l_i}),\ i = 1, \ldots, s,\ 1 \leq l_i \leq l_{i+1} \tag{3.11}$$

$$\mathrm{Ann}([g_i]) = 0,\ i = s+1, \ldots, s+r. \tag{3.12}$$

と表せる。よってこの表示において、$d_i = deg\ g_i$ とすると、生成元 $[g_i]$ は時刻 d_i の単体複体 K^{d_i} で新たに発生するホモロジー類を示す。g_i に対して、時刻 d_i は発生した時刻を示し、l_i は存続区間を表すことになる。

パーシステントホモロジー群に対して、

$$I_i = \begin{cases} [d_i, d_i + l_i), & i = 1, \ldots, s, \\ [d_i, \Theta], & i = s+1, \ldots, s+r \end{cases} \tag{3.13}$$

をパーシステント区間と呼ぶ。ここで Θ はフィルトレーションの飽和時刻である。また d_i をパーシステント区間 I_i の発生時刻、$d_i + l_i$ を消滅時刻と呼ぶ。

また、パーシステント区間 I_i に対して、$I_i(b)$ で区間の下限 (birth),$I_i(d)$ で区間の上限 (death) を表すことにする。

パーシステントホモロジー群に対して

$$PH_k(\mathbb{K}) = \{(I_i(b), I_i(d)) \in \mathbb{R}^2 \mid i = 1, \ldots, s+r\}. \tag{3.14}$$

を k 次パーシステント図と定める。

ここで $l_i \geq 1$ かつ $d_i \leq \Theta$ より、パーシステント図 $PH_k(\mathbb{K})$ 内すべての点は対角線上より上側にくる。例として、図 3.1 にロボカップサッカーシミュレーション 2D のログからサンプルした 2 次元サンプル点データ、図 3.2 にサンプル点データのパーシステント図を示す。パーシステント図の点はホモロジー類を示し、横軸は (birth)、縦軸は (death) を示す。

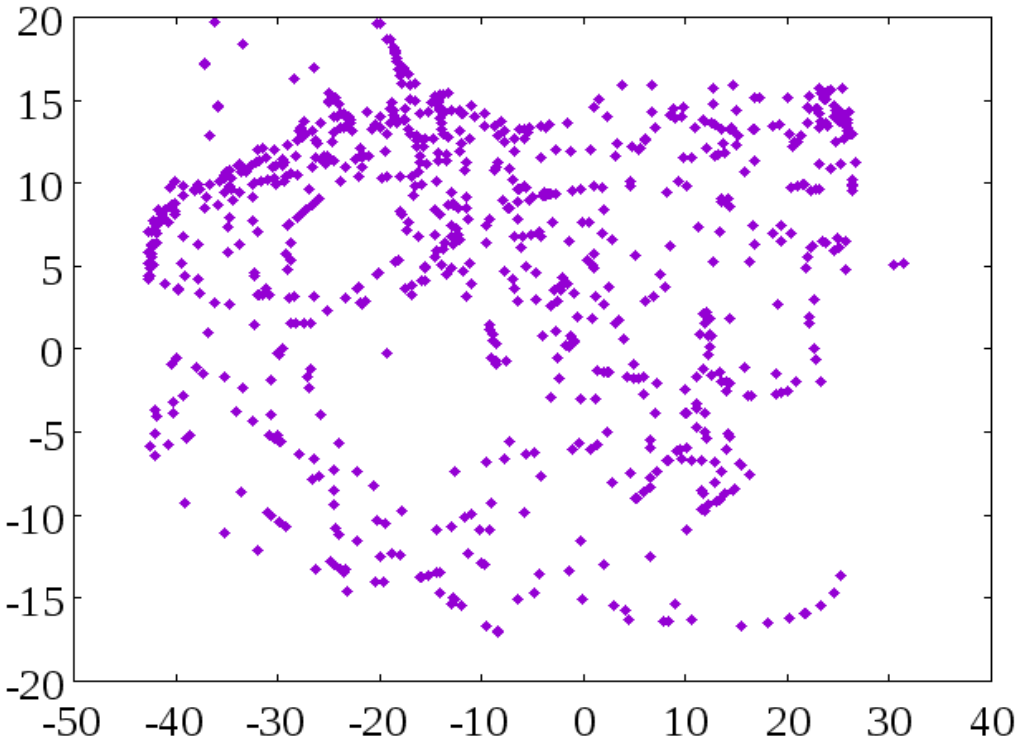

図 3.1: サッカー行動からサンプルしたサンプル点データ

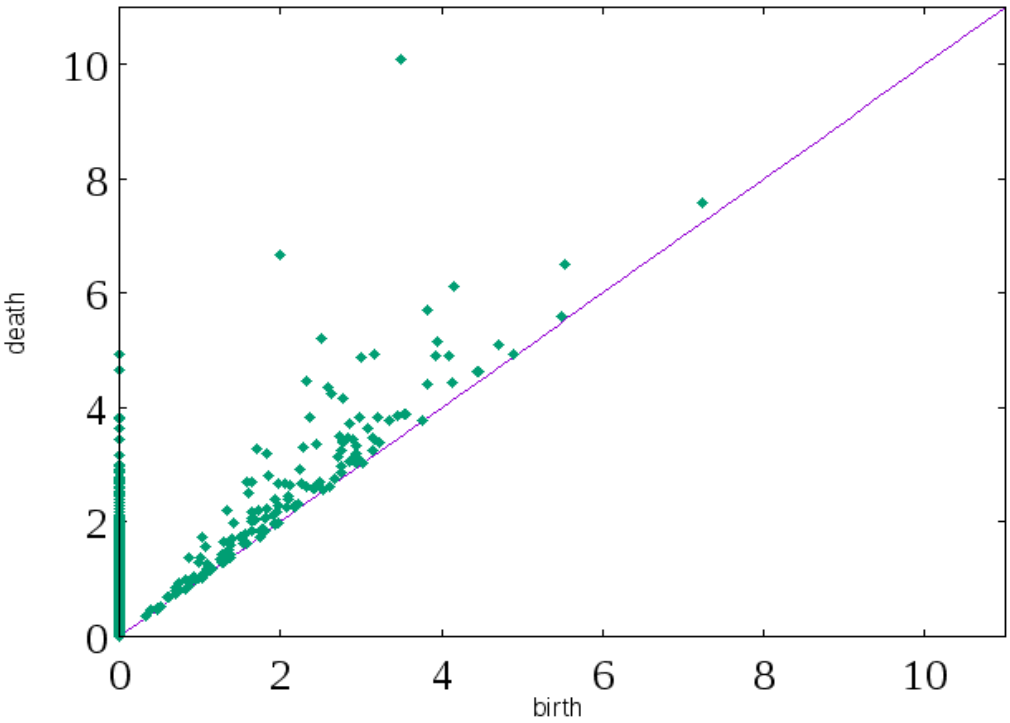

図 3.2: サンプル点データのパーシステント図

第4章　パーシステントホモロジーによる多次元ファジィ集合生成手法

4.1　提案手法

パーシステントホモロジーを用いて、幾何学的構造を保存するための閾値とサンプル密度を計算する閾値を導出する。

サンプル点データの構造を同定するためにパーシステントホモロジーを用いて計算を行う。パーシステントホモロジーの計算にはRipser[13][14]を用いた。

パーシステントホモロジーで生成されたホモロジー群の中には、誕生したタイムステップと消滅したタイムステップが近いものがある、短いタイムステップでしか生存しなかったホモロジー群はデータの幾何学構造よりもノイズである可能性が高い。よって、ノイズの影響を少なくするために重み付けを行う。

重み付けに関しては、先行研究[15]の手法を用いる。重み関数

$$
\begin{aligned}
&w(x) = \arctan(C\mathrm{Pers}(x)^p), \\
&(C, p > 0), \\
&\mathrm{Pers}(x) = d - b \text{ for } x \in \{(d, b) \in \mathbb{R}^2 | b \leq d\}.
\end{aligned}
\tag{4.1}
$$

を用いて、重み付けを行う。

重み関数はパラメータC, pを調整することで、パーシステントホモロジーへの影響を調整することができる。先行研究より、$p = 5$、$C = (\mathrm{midium}\{\mathrm{Pers}(x_i) | x_i \in D\})^{-p}$と設定した。

例として、図4.1にロボカップサッカーシミュレーション2Dのログからサンプルした2次元サンプル点データ、図4.2にサンプル点データのパーシステント図、図4.3に重みを付けたサンプル点データのパーシステント図を示す。

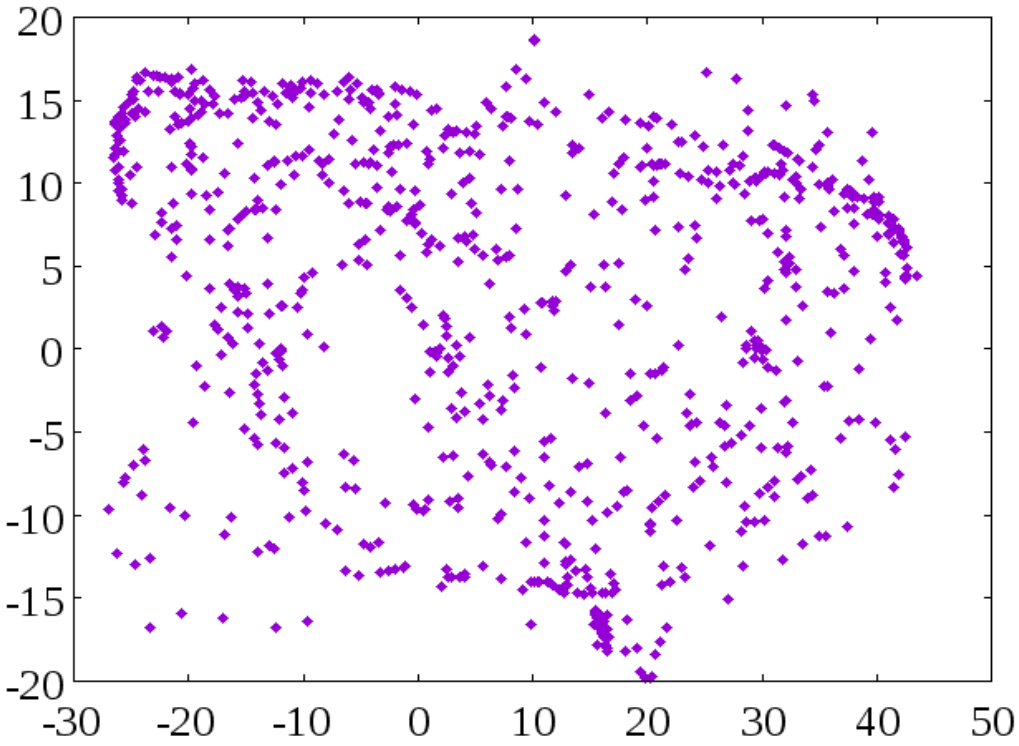

図 4.1: サッカー行動からサンプルしたサンプル点データ

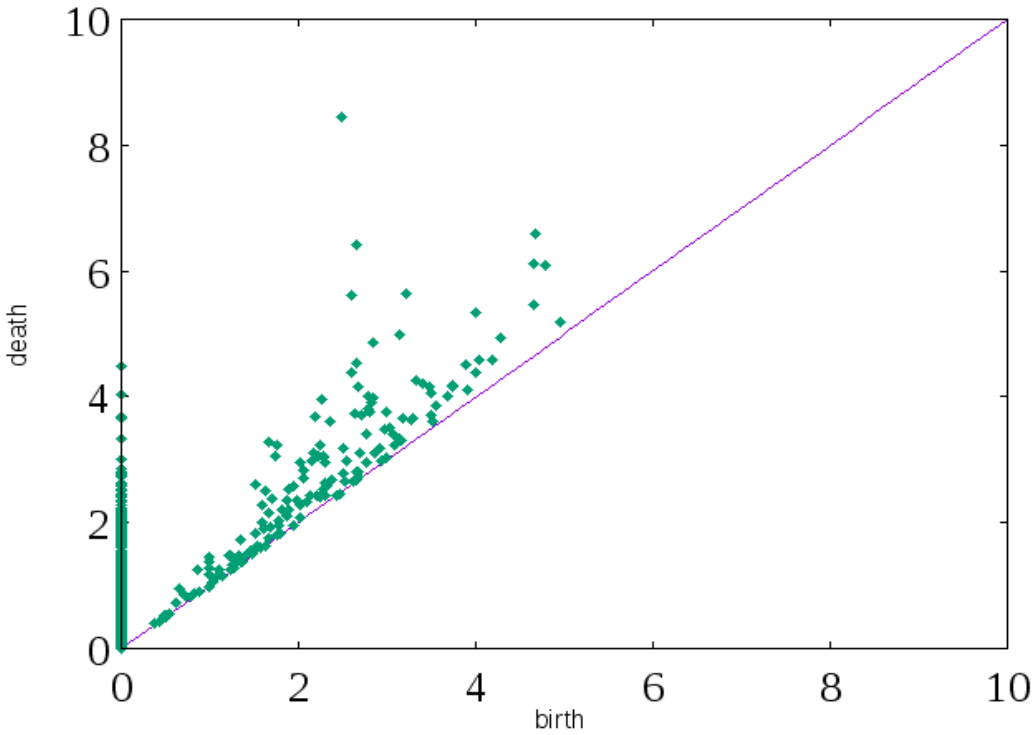

図 4.2: サンプル点データのパーシステント図

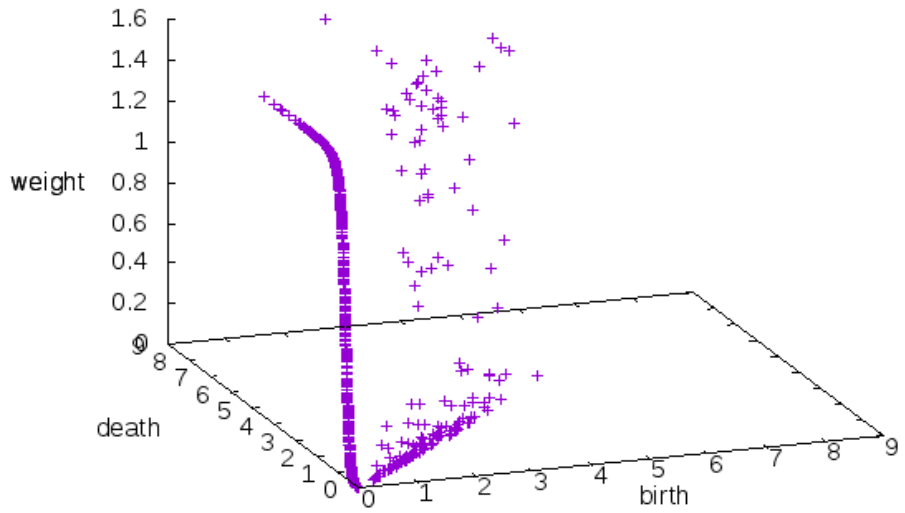

図 4.3: サンプル点データの重み付きパーシステント 図

重みを 付けられたパーシステント 図に対して、0 次元ホモロジー群を 除き、b_i と d_i の中点の重み付け平均を 計算し、この平均値を 近傍閾値 ϵ_M とし、d_i の重み付け 平均を 計算し、この平均値を 近傍閾値 ϵ_F とする。$w(x_i)$ は式 4.1 によって 求める。

$$\epsilon_M = \frac{\sum_{i=1}^{i \leq N} w(x_i)(b_i + d_i)}{2 \sum_{i=1}^{i \leq N} w(x_i)}. \tag{4.2}$$

$$\epsilon_F = \frac{\sum_{i=1}^{i \leq N} w(x_i) d_i}{\sum_{i=1}^{i \leq N} w(x_i)}. \tag{4.3}$$

4.2 比較実験

4.2.1 実験設定

多次元ファジィ集合生成方法を用いて、二つの閾値に対し、提案手法と手動で求めた場合の比較実験を行った。

実験にはロボカップサッカーシミュレーション 2D と呼ばれる 2D サッカーシミュレーションのログデータを使用した。

- ある試合 6000 点,46 次元のデータから、ボールの移動ログデータ,1000 点をランダムサンプリングした 2 次元データ $(x_b(t), y_b(t))$
- ある試合 6000 点,46 次元のデータから、オフェンス 2 人の移動ログデータ,400 点をランダムサンプリングした 4 次元データ $(x_1(t), y_1(t), x_2(t), y_2(t))$

を用いた。

図 4.4 にボールの移動ログデータ,1000 点をランダムサンプリングしたデータを示す。

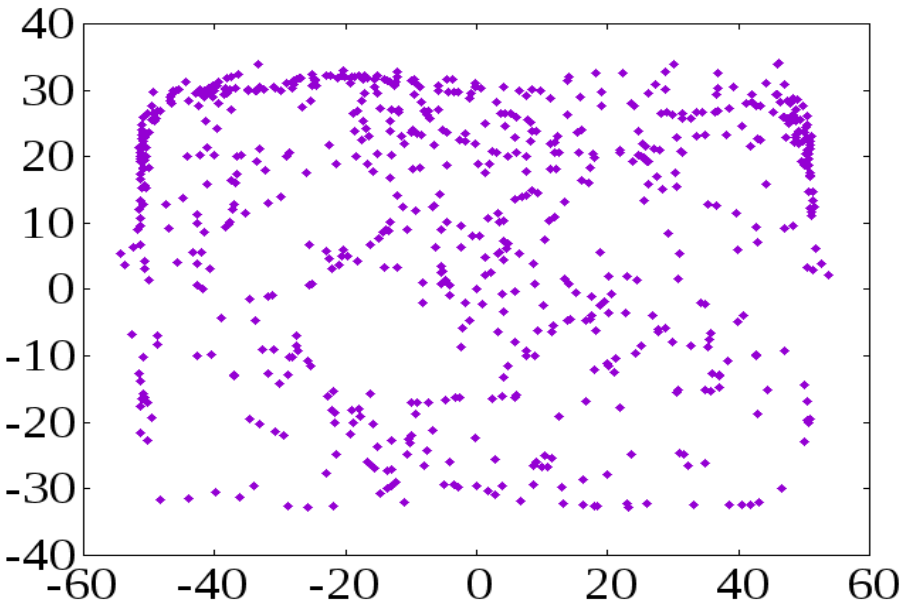

図 4.4: サンプル点データ

閾値 ϵ_M は、手動で K_M=3,9 と設定したものと、提案手法で求めたもの、閾値 ϵ_F は、手動で K_F=9,27 と設定したものと、提案手法で求めたものを用いた。二つの閾値のそれぞれの組み合わせで多次元ファジィ集合を生成し、比較検討を行う。

4.2.2 実験結果

ある試合における、ボールの移動ログデータから 1000 点ランダムサンプリングした 2 次元データからそれぞれの閾値の組み合わせによって生成した多次元ファジィ集合を図 4.5～図 4.13

に示す。また、それぞれの多次元ファジィ集合のネットワーク構造のパス数を表 4.1 に、メンバーシップ値の分散を表 4.2 に示す。

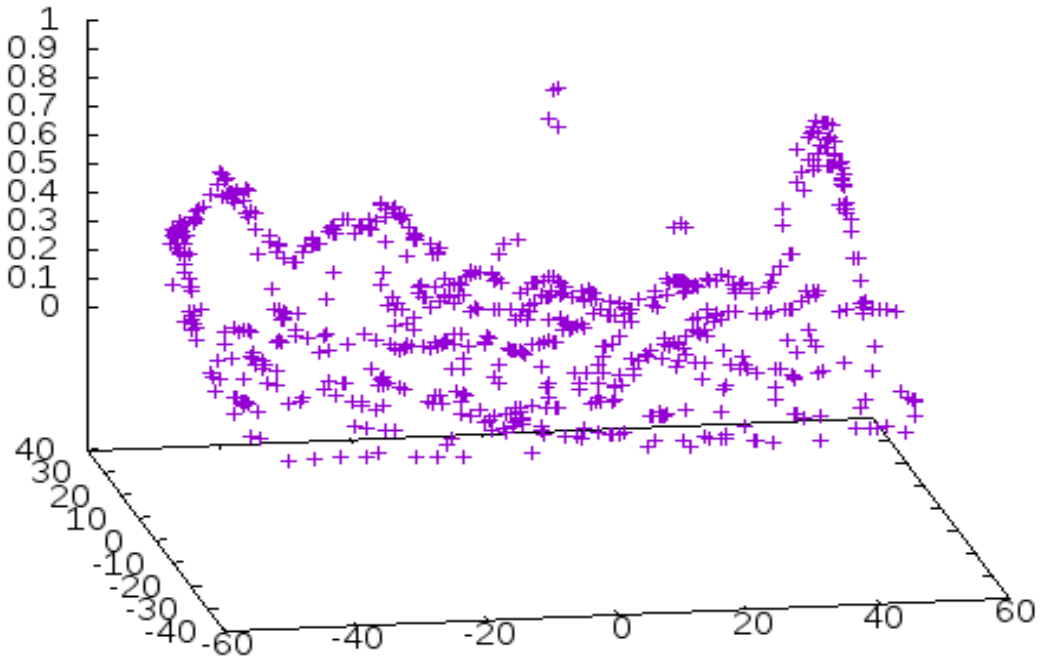

図 4.5: $K_m = 3, K_f = 9$

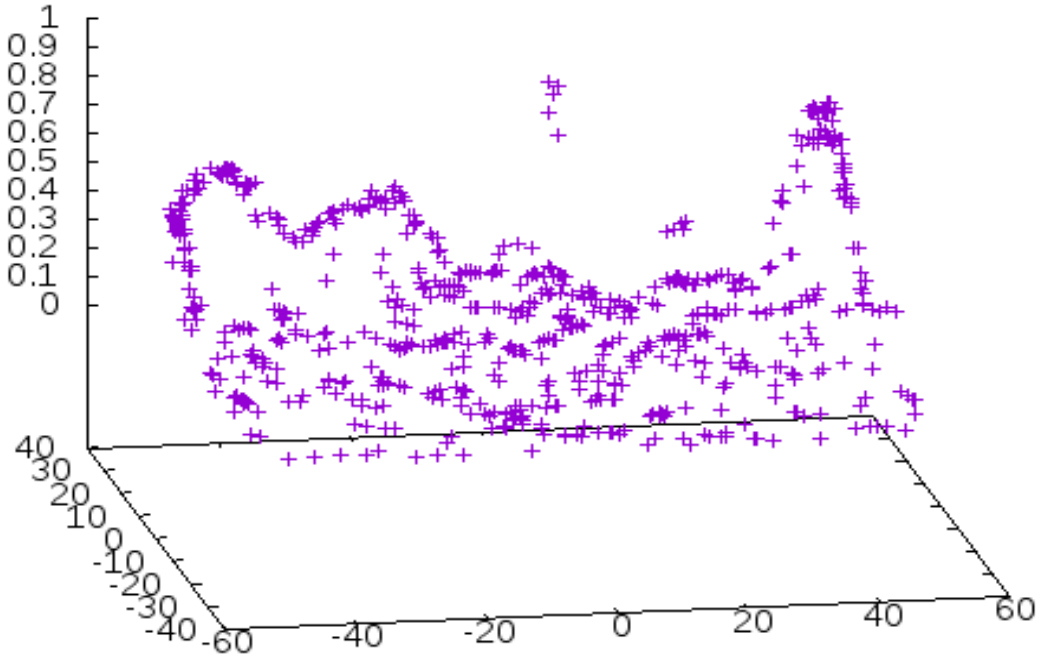

図 4.6: $K_m = 3, \epsilon_F$: 提案手法

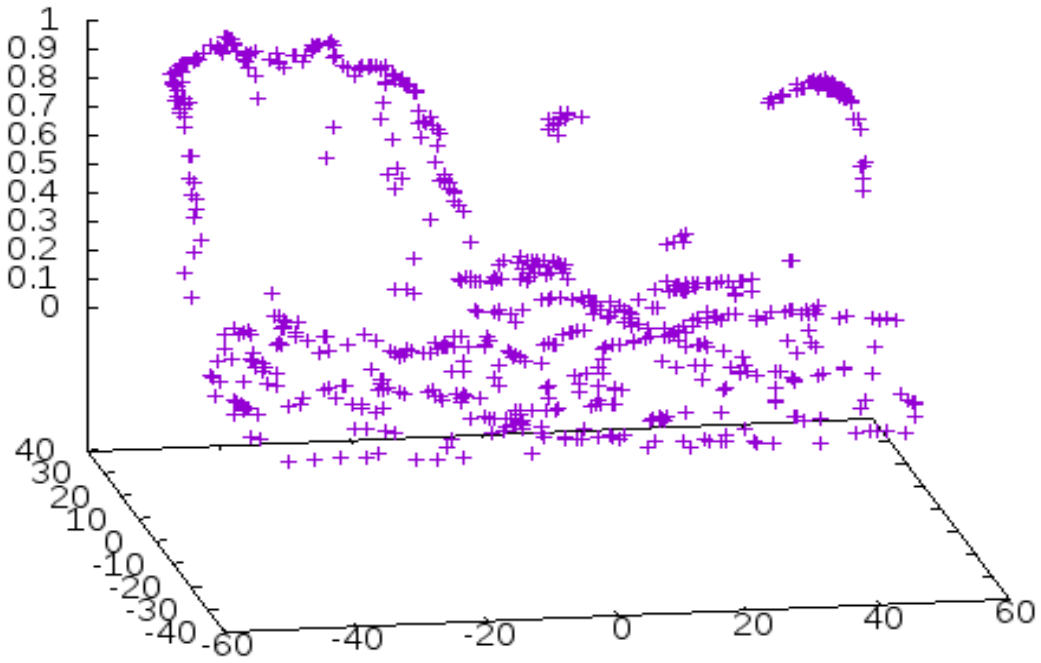

図 4.7: $K_m = 3$,$K_f = 27$

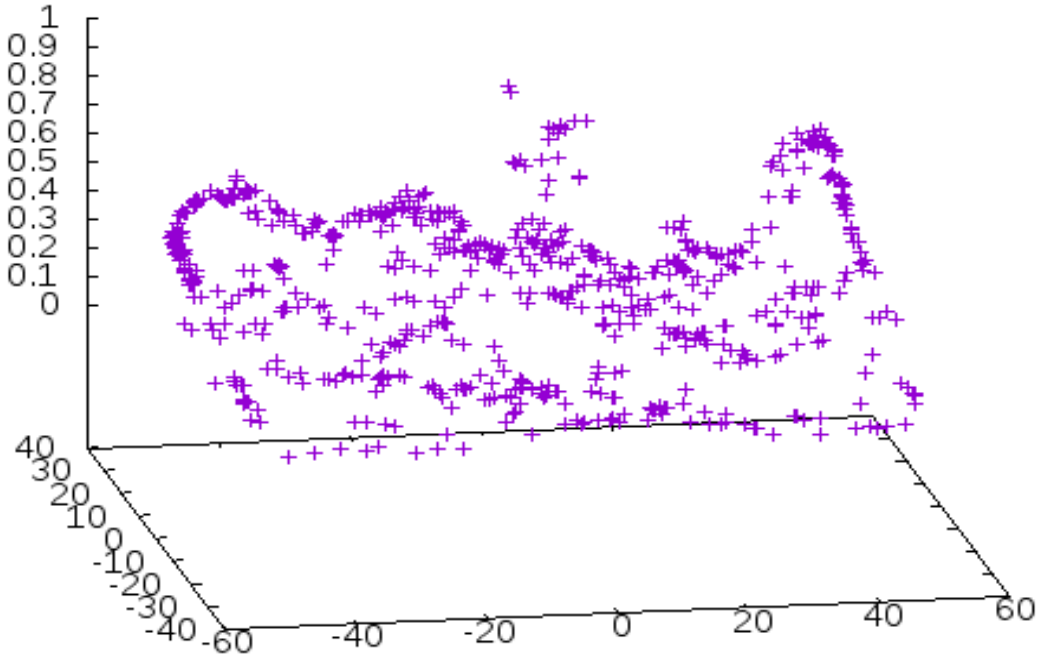

図 4.8: ϵ_M : 提案手法,$K_f = 9$

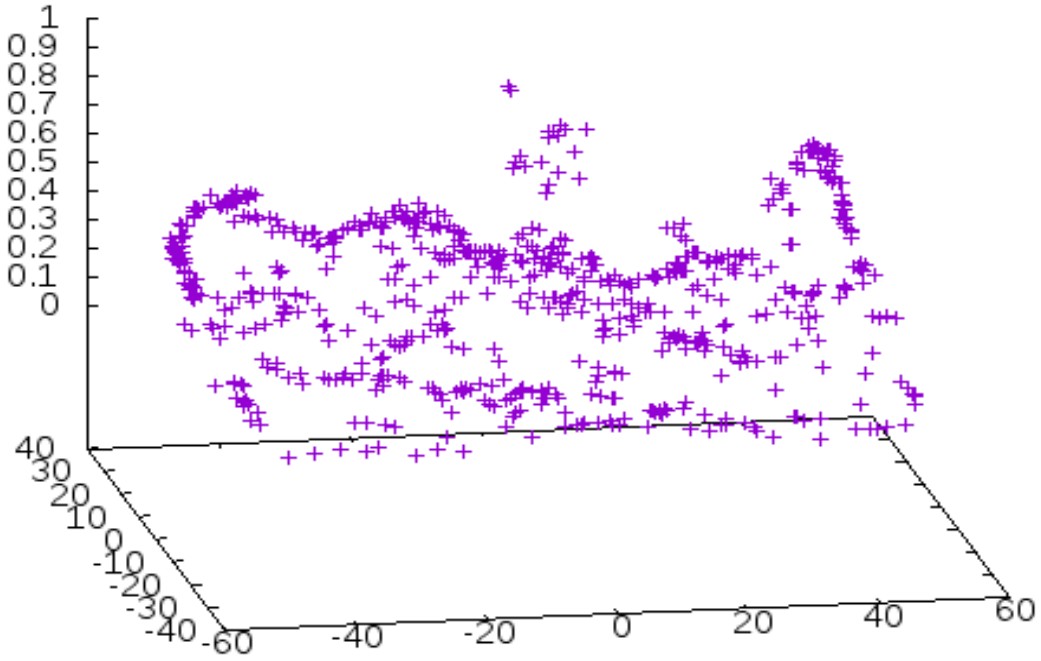

図 4.9: ϵ_M : 提案手法,ϵ_F : 提案手法

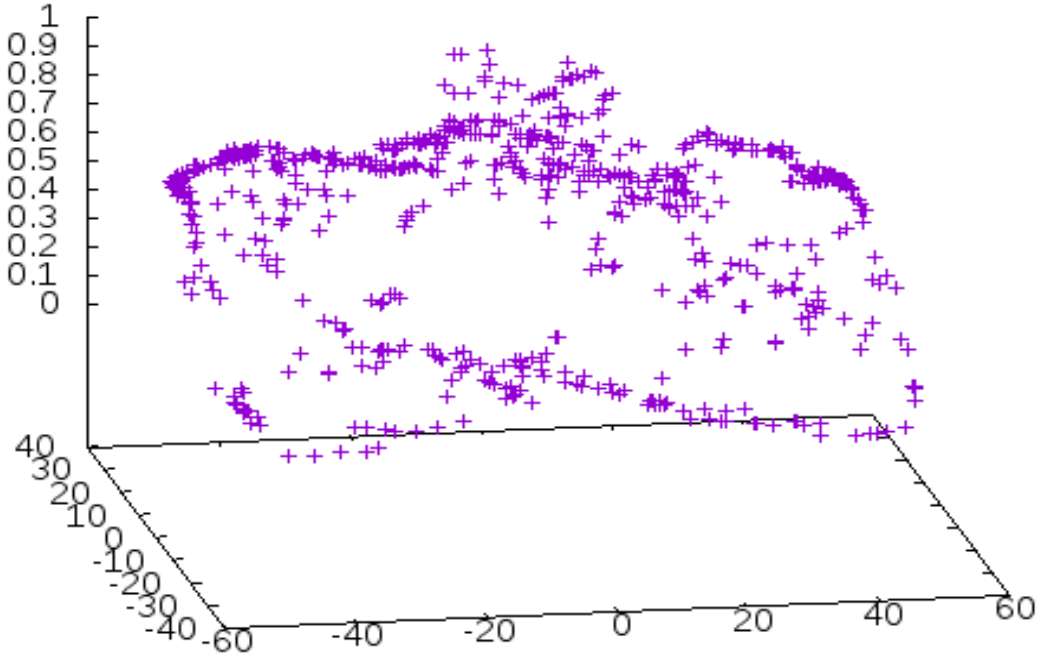

図 4.10: ϵ_M : 提案手法、$K_f = 27$

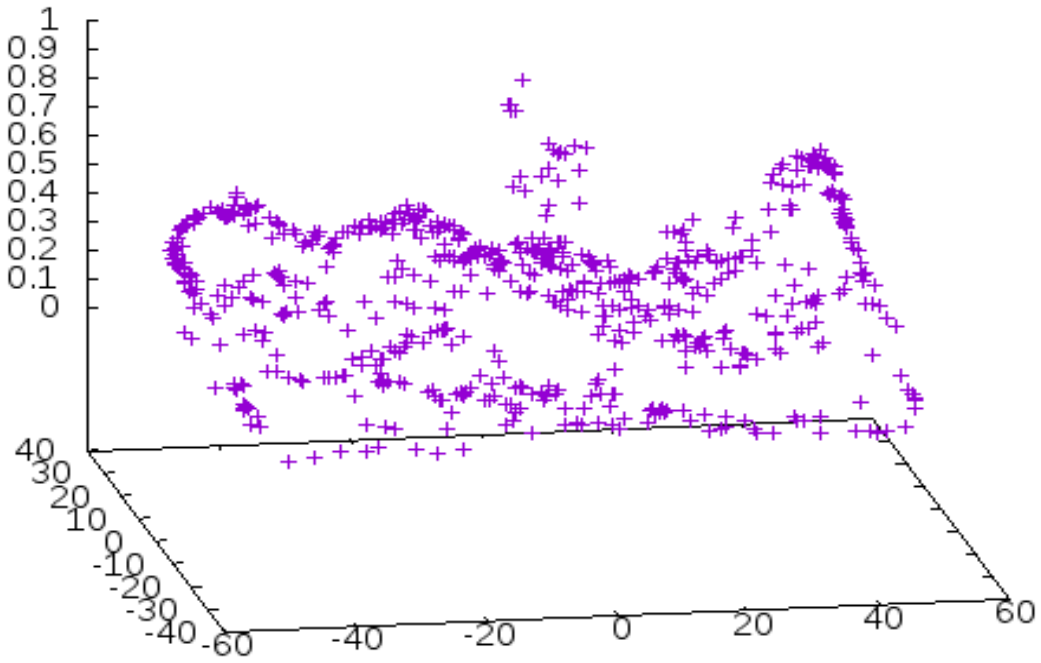

図 4.11: $K_m = 9$,$K_f = 9$

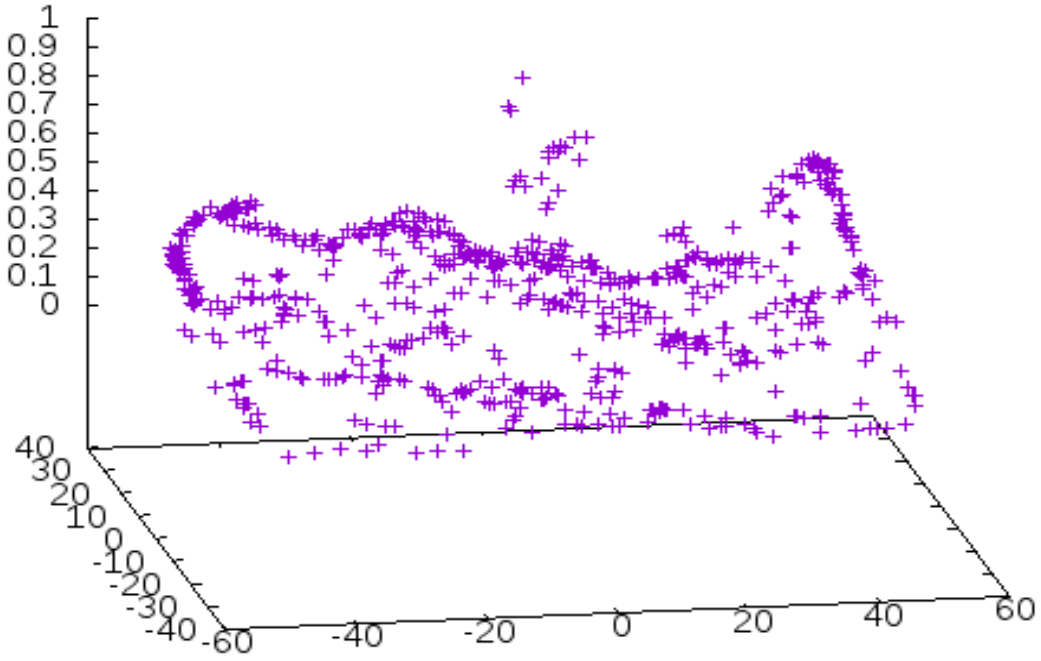

図 4.12: $K_m = 9$,ϵ_F : 提案手法

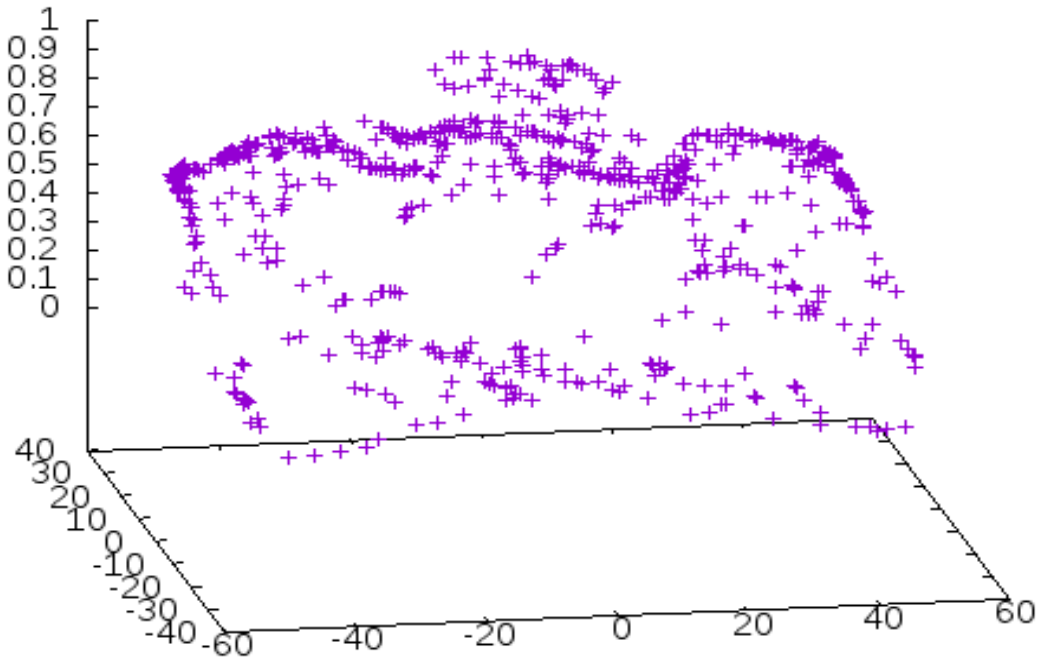

図 4.13: $K_m = 9$,$K_f = 27$

表 4.1: ネットワーク構造のパス数

ϵ_M	パス数
$K_m = 3$	6287
提案手法	13463
$K_m = 9$	18395

表 4.2: メンバーシップ値の分散

ϵ_M \ ϵ_F	$K_f = 9$	提案手法	$K_f = 27$
$K_m = 3$	0.059381	0.067638	0.124856
提案手法	0.045772	0.046139	0.051188
$K_m = 9$	0.036028	0.038696	0.047227

ある試合における、オフェンス 2 人の移動ログデータから 400 点ランダムサンプリングした 4 次元データからそれぞれの閾値の組み合わせによって生成した多次元ファジィ集合のネットワーク構造のパス数を表 4.3 に、メンバーシップ値の分散を表 4.4 に示す。

表 4.3: ネットワーク構造のパス数

ϵ_M	パス数
$K_m = 3$	1794
提案手法	3339
$K_m = 9$	15399

表 4.4: メンバーシップ値の分散

ϵ_M \ ϵ_F	$K_f = 9$	提案手法	$K_f = 27$
$K_m = 3$	0.079824	0.047352	0.073158
提案手法	0.058190	0.042661	0.048680
$K_m = 9$	0.035861	0.042289	0.016258

4.2.3 考察

2 次元データにおいて、この図形では上側に点が偏り、左上と右上に穴が開いている分布だと確認できる、二つの閾値をパーシステントホモロジーで導出したものが、そのデータの分布をよく表すことができている。メンバーシップ値の分散を計算した場合、最も分散が大きくなるような閾値はパーシステントホモロジーから導出した閾値とは異なっている。これは、メンバーシップ値が 1 に近い値と 0 に近い値に二分しているような多次元ファジィ集合の分散が大きくなっているためである。ここから、分散が大きければデータのバリエーションが多くなるとは限らないことが確認できた。

4 次元データにおいて、$K_m = 3$ ではパス数が少ないことからデータの偏りが著しく、分散が大きくなっていると考えられる。提案手法が導出した閾値 ϵ_M はパス数は多く、分散の値も $K_m = 9$ よりも大きくなっていることから、他の値よりも良い多次元ファジィ集合が生成することができている。しかし、提案手法が導出した閾値 ϵ_F は、$K_f = 9, 27$ 両方よりも分散が小さくなってしまっており、他の閾値より偏った多次元ファジィ集合になっていると考えられる。

本研究では、0 次元ホモロジー群を除いて、パーシステント図の重み付き平均を計算している。これは、0 次元ホモロジー群がデータ点の数だけ生成され、birth が 0 となっており、重みを付けたとしても大きな影響を与えるため、平均値が非常に小さくなってしまうためである。しかし、0 次元ホモロジー群もデータの幾何学的特徴ではある。削除するのではなく、ホモロジー群の次元による重みの変更などを考える必要がある。

第5章　結論と今後の展望

5.1　結論

本研究により、データのパーシステント図を計算し、0次元ホモロジー群を除き重み付き平均をとることで閾値ϵ_M,ϵ_Fの指標となる値を式5.1,5.2のように提案することができた。ここでb_iはホモロジー群x_iにおける birth、d_iはホモロジー群x_iにおける death、$w(x_i)$は式5.3によって求めるホモロジー群x_iにおける重みである。

$$\epsilon_M = \frac{\sum_{i=1}^{i\leq N} w(x_i)(b_i + d_i)}{2\sum_{i=1}^{i\leq N} w(x_i)}. \tag{5.1}$$

$$\epsilon_F = \frac{\sum_{i=1}^{i\leq N} w(x_i)d_i}{\sum_{i=1}^{i\leq N} w(x_i)}. \tag{5.2}$$

$$\begin{aligned} &w(x) = \arctan(C\mathrm{Pers}(x)^p) \\ &(C, p > 0), \\ &\mathrm{Pers}(x) = d - b \text{ for } x \in \{(d, b) \in \mathbb{R}^2 | b \leq d\}. \end{aligned} \tag{5.3}$$

パーシステントホモロジーはデータの幾何学構造を同定することができ、多次元データでも用いることができるため、多次元ファジィ集合へ応用させることができた。また、重みを付けているため birth と death が近いノイズであるホモロジー群の影響を小さくすることができている。0次元ホモロジー群を除いているため、データの連結成分に関する幾何学構造を同定することはできていないが、他次元のホモロジー群による幾何学構造を同定することで、幾何学構造を保存するような閾値を提案できた。

人が視認できない多次元空間において、幾何学構造を用いることで多次元ファジィ集合を編集するための検討材料の一つにすることができたと言える。

5.2　今後の課題

本研究の課題はパーシステントホモロジーからよりよい閾値を導出することである。平均を取った場合、0次元ホモロジー群の birth が多いため値が小さくなる傾向がある。本研究では0

次元ホモロジー群を除いて計算したが、0 次元ホモロジー群も幾何学構造であるために、ホモロジー群の次元によって重みを変更するなど方法を検討する必要がある。

サンプル密度を計算するための閾値は、本研究の手法では幾何学構造を保存するための閾値と近くなり、幾何学構造を保存した意味がなくなってしまうため、他の計算手法も検討する必要がある。

また、本研究によって提案した閾値は人が編集する際の指標の一つであり、より人が多次元データを理解するために他の検討材料を提案することが課題である。統計量を用いることや、人が理解できるよう次元を削減したデータを用いることが必要になる。

そして、システム面から考え、システムにおいて使いやすいためにメンバーシップ値の分散が最も大きくなる閾値を取るという考え方もある。様々な点から検討できるようにすることが一番の課題である。

参考文献

[1] Lotfi A Zadeh. The concept of a linguistic variable and its application to approximate reasoning—i. *Information sciences*, Vol. 8, No. 3, pp. 199–249, 1975.

[2] Lofti A Zadeh. Information and control. *Fuzzy sets*, Vol. 8, No. 3, pp. 338–353, 1965.

[3] Takaki Okuyama, Nobuhiko Kawashiro, and Junji Nishino. Humanoid motion generation with fuzzy state knowledge. In *27th Fuzzy System Symposium*, pp. 175–178, 2011.

[4] Junji Nishino and Akihiro Kasuya. Gpgpu for human modelling with konohen fuzzy set. In *27th Fuzzy System Symposium*, p. to be appeared, 2011.

[5] Akihiro Kasuya and Junji Nishino. High-speed generation of multi-dimensional, fuzzy set with gpgpu. In *26th Fuzzy System Symposium*, pp. 1232–1235, 2010.

[6] Junji Nishino. Konohen fuzzy: A sample points based computational model for multi-dimensional fuzzy set. In *Granular Computing (GrC), 2014 IEEE International Conference on*, pp. 230–234. IEEE, 2014.

[7] Takashi Harada and Junji Nishino. Multi-dimensional fuzzy set identification using persistent homology. *International Fuzzy Systems Association*, 2017.

[8] 原田貴史, 西野順二. 多次元ファジィ集合生成におけるパーシステントホモロジーの応用. 第 33 回ファジィシステムシンポジウム, 2017.

[9] Yasuaki Hiraoka. *Protein Structure and Topology : Introduction to Persistent Homology*. Kyoritsu Shuppan, 2013.

[10] Mark Anthony Armstrong. *Groups and Symmetry*. Springer Japan, 1988.

[11] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. In *Discrete Comput Geom 33*, pp. 249–274, 2005.

[12] Herbert Edelsbrunner and John Harer. Persistent homology - a survey. In *Contemporary Mathematics Volume 453*, pp. 257–282, 2008.

[13] Ulrich Bauer. Ripser. `https://github.com/Ripser/ripser`, 2015–2016.

[14] Otter N, Porter MA, Tillmann U, Grindrod P, and Harrington HA. A roadmap for the computation of persistent homology. `http://arxiv.org/abs/1506.08903`, 2015.

[15] Genki Kusano, Kenji Fukumizu, and Yasuaki Hiraoka. Persistence weighted gaussian kernel for topological data analysis. *International Conference on Machine Learning*, pp. 2004–2013, 2016.

謝辞

本研究に際して，日頃より様々なご指導を頂いた西野順二助教に心より御礼申し上げます．また，様々な手続きでお世話になった主任指導教員の緒方秀教教授，合同ゲームゼミで貴重なご意見を頂いた保木邦仁准教授そして研究生活の中で多くの助言を頂いた研究室の皆様に深く感謝致します．