データ解析技術による意思決定支援
Support for decision making by data analysis

1012154　杉澤智己　Tomoki Sugisawa

1　背景

近年、大量のビッグデータにおける解析が重要視されてきている。本プロジェクトでは勘に頼った意思決定ではなく、データ解析技術による意思決定支援をしてくれるシステムの作成をテーマに活動を行った。そして我々は人数が多かったため、テーマとグループを 2 つに分けシュートフォーム改善システムの作成を行うグループ A、コース決定支援システム作成を行うグループ B で活動を行った。

2　課題の設定と到達目標

システム作成を行うにあたって、メンバー全員でテーマ決定の議論を行った。テーマ決定する際にメンバー全員にテーマを考えてきてもらい、その中でシステムにおいてどのような入力に対してどのような出力をしてくれるか、システムを作成するにあたって、対象のユーザー、この 2 つが重要になってくるためその点に注意しなければならなかった。また入力と出力に驚きがあるシステムが良いという担当教員からのアドバイスなどもあり、議論に挙がった様々なテーマを絞るため全員で投票制にし、票が多い順に最終的に 2 つに絞った。最終的に残ったのはキネクトを使用し、動作解析、本学のコース決定支援の 2 つのテーマが残り、それぞれのテーマでシステム作成を行うこととなった。そしてそれぞれの各グループでの活動に移った。グループ A では動作解析によるバスケットボールのシュートフォームの改善のコメントやシュートの入る確率などを出力するシステムの開発、グループ B では本学のコースを決定する際の支援してくれるシステムの作成を目標に活動を行った。

3　課題解決のプロセスとその結果

3.1　機械学習-R 言語の学習

まず初めに我々はプロジェクト全体で意思決定支援を行うために必要な機械学習を用いた。機械学習とはコンピュータに人のような学習能力を獲得させるための技術の総称である。本プロジェクトのテーマは意思決定支援システムの作成であるため、自然言語処理である R 言語も用いた。しかし、我々は R 言語、機械学習について全くの無知であったことから初めに学この二つの学習を行った。web 上にある函館の過去 364 日分の気象データを使用し、天気予報判別システムを作成した。システムに学習能力を獲得させるため気象データを使用する。この気象データは一日の気温、湿度、風速、気圧、天気の 5 つであり、これが 364 日分のデータとなっている。ここで天気予報判別システムとはその日その日の気温、湿度、風速、気圧によっての天気 (訓練データ) から傾向や、特徴を見出し、これとはまた異なる新たな日のデータ (テストデータ) の気温、湿度、風速、気圧ではどのような天気になるか予想し、判別してくれるというものである。したがって気温、湿度、風速、気圧のデータを入力したときにそのデータの場合の天気を出力してくれるシステムということである。そのためシステム作成においてどのような特徴や傾向を見出すための訓練データを扱うかが重要になってくる。次にこのデータと R 言語を使いシステム作成を行った。システム作成の方法としては web 上にある CSV ファイルの気象データをダウンロードし、R 言語上で読み込み、特徴や傾向を解析するというものである。この解析には線形判別分析 (LDA) とサポートベクターマシーン (SVM) の 2 つの判別手法を使用し、解析を行った。線形判別分析とは条件付き確率 $p(x|y)$、つまり出力が y のときに特徴が x となる確率を考え、確率について正規分布モデルを用いてるものである。この線形判別分析を R 言語で使用し、先ほどの訓練データを解析することによってシステム完成である。しかし、これではテストデータがなくシステムの評価ができないため、訓練データを半分にし、半分を訓練データ、半分をテストデータとし、新たにシステム作成を行った。テストデータによる評価を行ったところ、線

形判別分析による手法では正答率約 81 %という結果が得られた。

3.2　システム作成　～グループ A～

我々がよく行うスポーツにとって、動作のフォームとはとても大事なものである。しかし、自分が理想としているフォームがはたして良いフォームといえるものだろうか。フォームを良いものかまた悪いものかと判断しするためにはそれなりの経験者、専門家が必要になってくる。また、どこにでもそういったあった存在があるとは限らない。そのため、スポーツにおけるフォームを撮影し、そのフォームの改善をしてくれるシステムの開発を目標に活動を行った。しかしスポーツとっても抽象的であり、様々スポーツのあらゆるフォームを解析するのはとても困難であり、扱うデータ量も仕事量もとても多くなってしまう。そのため、スポーツを 1 つに絞り、我々グループ A はバスケットボールのシュートフォーム改善システム作成について活動を行った。である。

3.2.1　画像解析の学習

まず初めに機械学習と R 言語の学習を兼ねて画像解析の学習を行った。ここでは適当な猫と犬の画像を集め訓練データとしてこれらを扱い、猫の写真か犬の写真かは別してくれるシステム作成を行った。画像データは RGB(カラー) であるため、3 次元になっており、全ての数値を扱うとても莫大なデータとなってしまう。そのため、画像をグレースケールに変換することにより、データも 1 次元になり、削減するができた。しかし、まだこれではデータ量が多く、R 言語上で警告が発生してしまった。そこで、画像解析の方法の 1 つでもある、Haar-like 法を用いて行った。そうすることにより、画像領域ごとの明暗さを訓練データとして扱うためデータ削減することが可能になりこの問題は解決することができた。

3.2.2　システム実装

画像解析でのシステムを動作解析で行うものとしたため、キネクトを用いた。まずキネクトを理解するために、サンプルプログラムを実際に起動しどのようなことができるか学習した。そしてキネクトでは搭載されている赤外線センサーを使用し、骨格取得 (頭、首、右肩、左肩、右肘、左肘、右手首、左手首、右手、左手、脊髄、腰、右臀部、左臀部、左膝、右膝、右足、左足) が可能であり、この骨格の座標の取得も行える。そのため、前述した天気用予報での学習を活かし、取得した骨格の座標を CSV ファイルへ書き込み、解析する方法に決定した。この方法を実現するため我々は開発環境である visual studio C ＃を使用し、wpf アプリケーションでのフォームアプリケーションの作成を行った。このフォームアプリケーションを使用しキネクトを操作し、CSV ファイルへの書き込みを行うというものである。しかし、キネクトで骨格の座標をただ取得する分けにはいけないため、タイミングや時間などの問題点の解決をするべく議論した。最終的に我々が出した取得方法は、フォームアプリケーション上で開始ボタンを設け、開始ボタンを押した地点から約 5 秒間 (20 フレーム) 座標を取り続けるというものである。CSV ファイルへ書き込み、解析する際にデータ数が異なってしまうといけないという問題点もあったためこの方法で決定した。また、プログラム上で骨格取得した数値をファイル書き込みへ順番とキネクトの座標取得の順番が合わないと数値も CSV ファイル入った時点で列や行が合わなくなってしまうため、キネクトの骨格取得の順番を調べた。キネクトによる骨格の順番は腰→脊髄→首→頭→左肩→左肘→左手首→左手→右肩→右肘→手首→右手→左臀部→左膝→左足首→左足→右臀部→右膝→右足首→右足となっている。次に解析方法の検討を行った。まず最初に検討したのがデータ数であった。今回扱うデータは 1 人 20 コマであり、必要な分だけの取得する骨格座標 (頭、首、右肩、左肩、右肘、左肘、右手首、左手首、右手、左手、腰、左膝、右膝、右足、左足) の 15 点、3 次元空間の座標を取得するため x 軸、y 軸、z 軸の 3 次元、つまり一人あたりのデータがとても多く解析するには複雑になってしまう。さらに取得した骨格の座標数値データが正しく取得されているかいないかのデバックをする必要がある。そのため、解析の方法として 1 人当たりのシュートフォームを打ってもらう時間、約 5 秒間の間に取得できる座標の中からある代表の座標を取り出し、扱うデータの量を減らす方法を採用した。また、代表の点はシュートするにあたって重要な部分であると考えた、肘、膝、脇の 3 点の角度である。シュートする際に腕の力と身体全体の上下運動をを行うための膝、ゴールに対しまっすぐに腕が向

いているかを判断するためにこの 3 点にした。デバックの問題はアニメーションを作成し、シュートフォームを骨格の取得した数値を基に点をプロットし、20 コマを順に表示していくことにより、問題を解決することができた。次にこのアニメーションを使用して予備実験を行ったところ、キネクトから取得した座標を用いて再生しているのだが、頻繁にシュートモーションのアニメーションとしておかしいと感じられるときがあり、何回も撮り直し、比較的綺麗にアニメーションが再生されたものをデータとして扱った。そしてシステム作成で我々はサポートベクターマシンを使用し、解析を行う。解析をするにあたって、次に予備実験の経験を生かし、データ収集を行った。我々が扱う訓練データは 50 人に 30 本シュートを打ってもらい入った本数を測定し、そのシュートの確率をラベルをとして扱った。これでシステムの完成である。しかし、これではただ単に、個人個人のシュートの確率を予想し、教えてくれるだけのものになってしまう。そこで、我々は前述したアニメーションを骨格座標の取得のデバックのためだけでなく、システムの出力画面として表示し、ユーザーにもシュートフォームが確認できるような HTML ページを作成することにした。また、アニメーションの表示だけでなく、シュートフォームの改善ということで、骨格の座標の数値を利用し、シュートフォームが前かがみになっていないか、などの改善コメントの表示もし、さらに求めた各角度での判別結果をパラメータとして表示することでこの問題の改善をすることができた。

3.3　システム作成　～グループ B～

グループ B はコース選択意思決定支援システムの作成について行った。本学では 2 年進級時にコースを 4 つの内から 1 つのコースを選択する。自分に合ったコースがどのコースか判断する際に、その選択の意思決定支援をするシステム開発を目的とし、活動を行った。

3.3.1　システム作成の学習

まず初めにシステム作成にあたって必要だと考えられる文章を単語単位で分解する形態素解析の学習を行った。この形態素解析をすることにより、複雑な「すももももももものうち」という文章も「すもも」「も」「もも」「も」「もも」「の」「うち」というように正しく名詞と助詞に分けることが可能であるためこの解析方法法を採用した。当初の段階としてまだ機械学習、R 言語になれていなかった我々は前述した天気予報のシステム作成の発展として形態素解析の学習を活かし、ニュース記事分類システムの作成を行った。これ読売新聞オンラインで公開されているニュース記事を使用し、記事を全 12 種類のジャンルごとに分類するものである。この分類には線形判別分析、最大エントロピー法、2 つの手法を使用し解析を行った結果、線形判別分析が 71 %、最大エントロピー法が 64 % であった。ここで両方の手法の結果を比較してみると両方とも共通して選挙の記事が誤って政治に分類されていることが分かった。

3.3.2　システム実装

次にニュース記事分類システムの作成で学んだことを基に次は、本題のコース決定支援システムの作成に取り掛かった。まず我々は各々のコースの特徴を知るべく各コース 20 人ずつ計 80 人、3 年生を対象としたアンケートを実施し調べた。そして特徴の結果として情報システムに属する人はプログラミングが好きで試験前に一夜漬けの勉強をしない、デザイン系の人は工作、おしゃれが好き、などの結果が得られた。しかし、アンケート内容について質問項目が 21 項目と多く、ユーザーにも同じ質問事項に答えてもらうには負担が多い考えた。この項目を減らすため我々は因子分析という手法を用いり、各質問事項の相関関係を見た。そこから重要度の高い上位 10 項目をデータ処理に使用することとしこの問題は解決することができた。そしてこの質問 10 項目のアンケートを 2,3,4 年生 230 人ほどに行いデータ収集を行った。アンケートを行う際に正確に答えてもらうためにチェック項目の『興味がない』から『興味を持っている』の順に 1 から 5 まで設け、途中でこの順番を逆にし、『興味を持っている』から『興味が無い』の順に 1 から 5 にするなどの工夫を施した。自由記述の項目に「分かりづらい」、「答えづらい」などの意見が出たため、普通のアンケートにした。そしてこの収集したデータを基に判別するのだが今回も模擬演習と同様に線形判別分析、サポートベクターマシン、ランダムフォレストの 3 つの機械学習手法から検討した。実際に判別した結果サポートベクターマシンがより適した手法で約 86 %の確率で判別できたためこの手法を採用した。システムはこれで完

成したが、科学的手法により判断はできるがただそれだけのものとなってしまう。そのため我々は平成 20 年から 25 年、過去 6 年分の卒業研究のタイトルと、研究性の所属しているコース、コース紹介資料から、各コースの特徴、各コースで学べることをまとめた。そしてこれらをデータ結果とともに表示し、ユーザーに分かりやすくするために HTML ページを作成した。さらに視覚的にも面白味のあるデザインにするためデータ班が各コースのイメージカラーについても調査を行い、グラフに取り入れた。

3.3.3　システムの評価

実際に 1 年生に使用してもらい視覚的にわかりやすく、「使用が簡易」、「各コースに関する情報が幅広く見られたので参考になった」と意見をもらい改善することができた。

4　今後の課題

4.1　グループ A

グループ A では問題点が多くなってしまい、大きく分けて、キネクト、システムの評価、データ収集の問題が挙げられる。具体的にキネクトの撮影の際の問題では以下の点である。

- ボールがで隠れてしまっている部分の手の座標取得があいまいになっている
- 撮影範囲に一人しか映ってはいけない
- 座標取得の効率が悪い
- 背景や撮影する場所の環境で座標取得できない場合がある
- 服装の色によって誤差が生じる
- コマの区切りが大幅である

この 6 つがキネクトにおける改善点である。そのためキネクトの特徴や性能をさらに生かし、今後に役立てるためのと同時に、これらの問題点を解決するべくキネクトについての学習と visual studio C ＃の学習がキネクトにおける今後の課題である。システムにおいては訓練データを用いてのシステム評価は行ったが、実際にテストデータでの評価を行っておらず、また実際にこのシステムを使用し、本当にシュートフォームが改善されるのか実験を行っていない。そのため実際に何人で一定期間の間システムを使用し、シュートの精度やフォームが改善されるかなどの評価を行うことが必要である。さらにデータ収集の問題でデータ数をより増やし、システムの精度をさらに上げることが課題である。

4.2　グループ B

システムの改良を行い、公開用のサーバーを立ち上げ、システムを web 上で公開して利用できるようにすること、また、1 年生に頂いた意見で根拠を示してほしいというものがあったためシステムに説得性を持たせることが今後の課題である。データ取集の課題として、4 年生に対してのあまりアンケートを実施することができなかった。また、個人情報の問題により、各コースごとの就職先のデータを収集することができなかった。そのため、よりよい精度のシステムを目指すためにも今以上にデータ収集を行う必要がある。

参考文献

[1] 中村 薫 (なかむらかおる). 新里祐教 (にいさとひろたか). 鷲尾友人 (わしおともと). OpenNI 3D センサープログラミング. 秀和システム, 2013.

[2] Microsoft. Kinect for Windows http://www.microsoft.com/en-us/kinectforwindows/.

[3] 下畑光夫, 因子分析 http://www.slideshare.net/mitsuoshimohata/ss-25635356/.

[4] フリーソフトによるデータ解析・マイニング第 25 回, R と因子分析 http://www1.doshisha.ac.jp/~mjin/R/25/25.html/.

[5] Rとカーネル法・サポートベクターマシン http://www1.doshisha.ac.jp/~mjin/R/31/31.html./