Analiza wyników

# Informacje ogólne - Jagielski

Pierwszym krokiem, aby porównać wybrany zbiór metod testowych jest ich przeprowadzenie oraz zebranie wyników. Jednak same wyniki liczbowe nie przynoszą żadnej wiedzy na temat testów. Dopiero po ich analizie można zacząć wyciągać wnioski. W poniższym rozdziale przedstawione zostaną użyte metody porównawcze. Opisany zostanie również tok myślenia, kierujący autorami podczas analiz.

Przeprowadzone zostały trzy różne scenariusze testowe, w sposób opisany w poprzednim rozdziale. Dwie z nich można porównać w sposób bezpośredni, ponieważ korzystają z tej samej puli sekwencji wideo. Filmy te oceniane są w tej samej skali w obu scenariuszach. Trzecia metoda badawcza, polegająca na porównywaniu dwóch następujących po sobie filmów, oceniana była w innej skali. Co więcej dostarczała wiedzy o filmach nie w porównaniu ze wszystkimi dostępnymi, ale tylko w zestawieniu z wybranym jednym. W dalszej części rozdziału zaproponowano autorskie rozwiązanie pozwalające porównać tak zestawione wyniki.

W części badawczej pracy magisterskiej zbadano prawdziwość tezy wpływu doboru scenariusza testowego na otrzymane wyniki.

# Test t-Studenta - Jagielski

W ramach analizy wyników przeprowadzono test t-Studenta. Test ten służy do porównania dwóch grup. Analizowana została średnia z każdej grupy, a następnie wykonane obliczenia pomogły w podjęciu decyzji o zachowaniu hipotezy zerowej.

Hipoteza zerowa jest to hipoteza, która poddawana jest weryfikacji. Założono w niej, że różnica pomiędzy uzyskanymi wynikami badań wynosi zero. W omawianej analizie zawartej w pracy magisterskiej hipoteza zerowa w teście t-Studenta dotyczyła zerowej różnicy między wynikami scenariuszy testowych, w którym oceniano każdą przedstawioną sekwencję wideo tylko raz, według narzuconej kolejności, a tą gdzie osoba oceniająca mogła wybierać oraz powracać do obejrzanych już filmów.

Wyniki każdego z wymienionych scenariuszy stworzyły osobną grupę. Istotną kwestią jest fakt, że obie grupy były niezależne od siebie, co indukuje fakt, że obie próby były od siebie niezależne. Efekt ten uzyskano dzięki losowaniu kolejności zarówno przeprowadzanych scenariuszy jak i odtwarzanych sekwencji filmowych. Zdecydowano się użyć testu t-Studenta również ze względu na brak danych o wartości średniej i odchylenia standardowego w całej populacji.

[http://lap.umd.edu/psyc200/handouts/psyc200\_0812.pdf]

Wraz z opisem przeprowadzonych obliczeń wyjaśniane będą kolejne pojęcia. Następnie omówione zostaną otrzymane rezultaty wraz z wyciągniętymi wnioskami.

Pierwszym krokiem przeprowadzanej analizy było obliczenie średniej oceny każdej grupy, dla każdego filmu. W każdej z grup znajdowała się taka sama ilość osób badanych. Skorzystano ze wzoru:

- średnia ocena dla j-tej grupy  
 - ilość przebadanych osób w j-tej grupie  
 - ocena i-tej badanej osoby, należącej do j-tej grupy

Kolejno obliczono różnicę wartości każdej oceny dla wybranego filmu i średniej grupy dla tej samej sekwencji wideo. Obliczenie należało wykonać dla każdego filmu oraz każdej grupy. Skorzystano ze wzoru:

- odchylenie od średniej i-tej oceny w j-tej grupie  
- średnia ocena dla j-tej grupy  
 - ocena i-tej badanej osoby, należącej do j-tej grupy

Otrzymane w ten sposób wartości podniesiono drugiej potęgi. [OPISAC DEWIACJE, TO JEST SUMA DEWIACJI DLA GRUPY] Następnie dla każdego filmu obliczono ich sumę.

- suma kwadratów odchyleń od średniej j-tej grupy  
 - odchylenie od średniej i-tej oceny w j-tej grupie  
 - ilość przebadanych osób w j-tej grupie

Kolejnym etapem było określenie liczby stopni swobody (ang. *degrees of freedom*). [OPISAC CO TO SA STOPNIE SWOBODY]. W przypadku przeprowadzonych badań liczba stopni swobody grupy równa była ilości badanych osób pomniejszonych o jeden.

- liczba stopni swobody j-tej grupy  
 - ilość przebadanych osób w j-tej grupie

Następnie oszacowano [KORZYSTAJĄC Z NIEOBCIĄŻONEGO ESTYMATORA NAJWIĘKSZEJ WIARYGODNOŚCI] wariancję dzieląc sumę kwadratów odchyleń od średniej przez liczbę stopni swobody dla każdej grupy. [SPRAWDZIC CZY NA PEWNO WARIANCJĘ I CZY NA PEWNO SZACUJEMY]

[http://www.naukowiec.org/wiedza/statystyka/stopnie-swobody\_718.html]

- estymowana wariancja j-tej grupy  
 - suma kwadratów odchyleń od średniej j-tej grupy  
 - liczba stopni swobody j-tej grupy

- [???]  
 - estymowana wariancja j-tej grupy   
 - liczba stopni swobody j-tej grupy

- [???]  
 - [???]  
 - ilość przebadanych osób w j-tej grupie

t - t-statystyka [???]  
- średnia ocena dla j-tej grupy  
 - [???]

Ostatnim krokiem było odczytanie z tablic rozkładu t-Studenta wartości krytycznej. Aby odszukać pożądaną wartość potrzeba znać liczbę stopni swobody oraz poziom istotności. Wartość krytyczna to liczba konieczna do stwierdzenia, czy hipoteza zerowa może zostać odrzucona na podstawie otrzymanych wyników. Odbywa się to poprzeć porównanie jej z wynikiem testu t-Studenta, w przedstawionych równaniach opisanym jako wartość *t*. Na tym etapie należy również wyjaśnić znaczenie poziomu istotności. Jest to liczba oznaczająca prawdopodobieństwo błędu, który jest akceptowalny w przeprowadzanym badaniu. Przez błąd rozumiany jest błąd pierwszego rzędu (ang. *false positive*). Pojawia się on, gdy hipoteza zerowa zostaje odrzucona, pomimo faktu, że w rzeczywistości jest ona prawdziwa. Określenie poziomu istotności ciąży na badaczu. W przeprowadzonej analizie przyjęto, że prawdopodobieństwo popełnienia błędu pierwszego stopnia wynosi 0.05.

|  |  |
| --- | --- |
| Numer sekwencji | Wynik testu t-Studenta |
| Sekwencja 1 | 0,85 |
| Sekwencja 2 | 0,49 |
| Sekwencja 3 | 0,74 |
| Sekwencja 4 | 0,83 |
| Sekwencja 5 | 1,23 |
| Sekwencja 6 | 0,41 |
| Sekwencja 7 | 0,51 |
| Sekwencja 8 | 0,27 |
| Sekwencja 9 | 0,21 |
| Sekwencja 10 | 0,83 |
| Sekwencja 11 | 1,11 |
| Sekwencja 12 | 0,51 |
| Sekwencja 13 | 0,30 |
| Sekwencja 14 | 0,50 |
| Sekwencja 15 | 0,22 |
| Sekwencja 16 | 0,70 |
| Sekwencja 17 | 0,49 |
| Sekwencja 18 | 0,25 |
| Sekwencja 19 | 0,66 |
| Sekwencja 20 | 0,47 |
| Sekwencja 21 | 0,31 |
| Sekwencja 22 | 0 |
| Sekwencja 23 | 0,21 |

Tabela 1. Wyniki testu t-Studenta dla poszczególnych sekwencji wideo.

W wyniku przeprowadzenia testu t-Studenta otrzymano rezultaty przedstawione w tabeli 1. Przy zadanej liczbie stopni swobody wynoszącej 12 oraz poziomie istotności 0.05 z tabeli rozkładu t-Studenta odczytano wartość 2,1788. Oznacza ona wartość minimalną, którą musiałby osiągnąć test t-Studenta dla większości sekwencji filmowej aby móc rozważać odrzucenie hipotezy zerowej, na rzecz hipotezy alternatywnej. W przeprowadzonym teście żadna wartość nawet nie zbliżyła się do tego progu.

Głównym wnioskiem z przeprowadzonej analizy t-Studenta iż przeprowadzone badania mające na celu porównanie dwóch wybranych metod testowych są nieistotne statystycznie. Wywnioskowano również na podstawie przeprowadzonego testu, iż postawiona hipoteza zerowa jest prawdziwa. Oznacza to, że w kontekście testu t-Studenta dla otrzymanych wyników wybór scenariusza testowego z wcześniej wymienionych nie ma wpływu na otrzymane wyniki. Można więc stosować je wymiennie, bez obawy o wypaczenie wyników.

# Autorska metoda translacji wyników w skali porównawczej na skalę pięciostopniową - Jagielski

Jednym z problemów, które pojawiły się podczas części badawczej pracy magisterskiej była niezgodność skali, użytych podczas testów subiektywnych. Dwa scenariusze operowały na skali pięciostopniowej, w której użytkownik miał wyrazić swoja opinię na temat wyświetlonego właśnie filmu. Trzeci bazował na skali siedmiostopniowej, pozwalającej użytkownikowi na ocenę porównawczą dwóch następujących po sobie sekwencji filmowych. Dostarczał zbioru wartości równemu kolejnym liczbom całkowitym w przedziale domkniętym obustronnie od minus trzech do trzech.

Przeprowadzono próbę przygotowania heurystycznego algorytmu, pozwalającego na przetłumaczenie wyników otrzymanych w omówionej skali porównawczej na skalę pięciostopniową. Algorytm pozwala na porównanie wyników otrzymanych skalach dowolnego stopnia, zarówno dostarczających danych wejściowych jak i wyjściowych. Omówiony zostanie jednak na przykładzie skali użytych w pracy magisterskiej.

Jednym z ograniczeń algorytmu jest sposób przygotowania scenariusza testowego. Musi on zostać przeprowadzony za pomocą scenariusza oceny porównawczej. Dane otrzymane z takiego testu muszą odpowiadać na pytanie: „Jak oceniasz film w porównaniu do poprzedniego?”. Scenariusz musi także zawierać filmy z całego przedziału jakości, zaczynając od bardzo dobrej, kończąc na bardzo słabej. Podczas tłumaczenia zostaje wykonane założenie, że najlepsza i najgorsza ocena filmu przyjmuje oceny brzegowe. Możliwe jest ograniczenie translacji do kilku elementów skali, jednakże podczas pracy magisterskiej nie stworzono algorytmu, pozwalającego na predykcję subiektywnych ocen testera.

Kolejne porównania przeprowadzane w teście powinny dotyczyć małych różnic w jakościach filmów. Zbyt duża rozbieżność w porównywanych jakościach może doprowadzić do wypaczenia wyniku translacji. Równocześnie w teście powinna zostać użyte filmy o takiej samej jakości jak w scenariuszu pięciostopniowym.

Drugim ważnym ograniczeniem, dotyczącym wyboru sekwencji do konkretnych par jest nazwane przez autorów ograniczenie ścieżki. Jeśli wszystkie sekwencje filmowe użyte w scenariuszu testowym są reprezentowane jako wierzchołki grafu, a ich zestawienie w porównywanej parze jako krawędź tego grafu o wadze równej ocenie, to konieczne do spełnienia są warunki można zapisać jako:

* Graf nie posiada żadnych pętli. Oznacza to, że nie można znaleźć co najmniej trzech filmów, które są ze sobą wzajemnie porównane.
* Graf posiada liczbę krawędzi |E| równą ilości wierzchołków |V| pomniejszonej o jeden. Oznacza to, że na zasadzie porównań zestawić możne dwa dowolne filmy ze zbioru badanych.

Algorytm wykonuje się iteracyjnie, za każdym razem wprowadzając nowe dane dla nowej, przetłumaczonej skali. W pierwszym kroku wybierana jest para, która nie była jeszcze przetworzona przez algorytm. Jeśli jest to pierwsza wstępnie tłumaczona para, wybranemu filmowi przypisuje się wartość zero. Kolejny z pary otrzymuje wartość uzyskanej oceny. Jeśli jakaś para została już wstępnie przetłumaczona algorytm wyszukuje pośród pozostałych, nie przetworzonych przez algorytm takich, w której dokładnie jeden z filmów pojawił się w poprzedniej. Film z nowej pary, któremu została już przypisana wartość przez algorytm nie zmienia jej. Drugiemu z pary nadawana jest wartość równa pierwszemu, odpowiednio zmieniona o wartość oceny porównawczej.

Powyższe czynności wykonywane są tak długo, aż wszystkie pary zostaną przetworzone przez algorytm. W tym momencie oczekiwanym rezultatem działania algorytmu jest graf w postaci ścieżki, w którym kolejne węzły wzdłuż jego przebiegu posiadają posortowane rosnąco lub malejąco wartości.

Kolejnym elementem jest wyznaczenie przedziałów, które będą odpowiadać elementom skali pięciostopniowej. W tym celu obliczana jest suma wartości bezwzględnych maksymalnej i minimalnej oceny wystawionej przez algorytm w poprzednich krokach. Otrzymaną liczbę należy podzielić przez cztery, aby wyznaczyć odstępy między kolejnymi liczbami odpowiadającymi ocenom ze skali pięciostopniowej. Następnie należy odczytać oceny algorytmu z kolejnych wierzchołków i sprawdzić, od którego z otrzymanych krańców przedziałów dzieli ich najmniejsza wartość. Może się zdarzyć, że ocena przyjmie wartość dokładnie w połowie między dwoma elementami skali pięciostopniowej. W drodze wyjątku należ nadać w takim przypadku ocenę połowiczną, na przykład cztery i pół, pomimo braku takiej wartości w skali pięciostopniowej.

Poniżej przedstawiono kolejne kroki algorytmu na przykładzie danych otrzymanych podczas testów subiektywnych opisywanych w pracy magisterskiej. Niestety scenariusz nie został przygotowany z uwzględnieniem wszystkich opisanych wcześniej ograniczeń, dlatego przykład powinien być traktowany jako zobrazowanie kolejnych kroków algorytmu.

|  |  |  |
| --- | --- | --- |
| Film1 | Film2 | Ocena porównawcza |
| Puppies\_7000k | Puppies\_1500k | -3 |
| Puppies\_9000k | Puppies\_4000k | -1 |
| Puppies\_3000k | Puppies\_2000k | -1 |
| Chimei\_1500k | Chimei\_2000k | 0 |
| Puppies\_3000k | Chimei\_1500k | 0 |
| Puppies\_1500k | Chimei\_1500k | 0 |
| Puppies\_9000k | Puppies\_7000k | 0 |

Tabela 2. Zestawienie porównanych filmów wraz z oceną jednego z testerów.

Postępując według kolejnych kroków algorytmu otrzymano względną ocenę wszystkich filmów. W tabeli 3 przedstawiono filmy posortowane po wartości przypisanej przez algorytm.

|  |  |
| --- | --- |
| Film | Wartość przypisana przez algorytm |
| Puppies\_9000k | 0 |
| Puppies\_7000k | 0 |
| Puppies\_4000k | -1 |
| Puppies\_3000k | -3 |
| Puppies\_1500k | -3 |
| Chimei\_1500k | -3 |
| Chimei\_2000k | -3 |
| Puppies\_2000k | -4 |

Tabela 3. Zestawienie filmów oraz wartości przypisanym im przez algorytm.

Następnie obliczono wartości punktów, odpowiedzialnych za przetłumaczenie ocen między skalami. Dzieląc sumę bezwzględnych maksymalnych i minimalnych wartości otrzymano licznik badanego ułamka, wynoszący w tym przypadku 4. Mianownik określany jest jako ilość możliwych ocen skali wyjściowej pomniejszona o jeden, w tym przypadku również wynosząca 4. Iloraz tych dwóch liczb pozwolił poznać odstęp między kolejnymi wartościami tłumaczącymi. Obliczonym krokiem było 1.

|  |  |
| --- | --- |
| Ocena skali 5-stopniowej | Estymowana wartość |
| 5 | 0 |
| 4 | 1 |
| 3 | 2 |
| 2 | 3 |
| 1 | 4 |

Tabela 4. Zestawienie ocen skali 5-stopniowej oraz odpowiadających im estymowanych wartości.

Ostatecznie na zasadzie poszukiwania najmniejszej różnicy oceny zostały przydzielone do sekwencji filmowej.

|  |  |
| --- | --- |
| Film | Ocena w skali 5-stopniowej |
| Puppies\_9000k | 5 |
| Puppies\_7000k | 5 |
| Puppies\_4000k | 4 |
| Puppies\_3000k | 2 |
| Puppies\_1500k | 2 |
| Chimei\_1500k | 2 |
| Chimei\_2000k | 2 |
| Puppies\_2000k | 1 |

Tabela 5. Zestawienie filmów oraz wartości przypisanym im przez algorytm.

Należy zaznaczyć, że ta prosta metoda jest bardzo niedokładna i mocno uzależniona od przeprowadzonych porównań. Aby translacja była jak najbardziej dokładna porównania powinny być przeprowadzane w parach bardzo zbliżonych jakości. Przedstawiony powyżej przykład ma na celu przede wszystkim zobrazowanie przebiegu algorytmu. Ze względu na brak przygotowanego scenariusza spełniającego wymagania translacji, uzyskane dane obarczone są dużym błędem.

|  |  |  |  |
| --- | --- | --- | --- |
| Film | Średnia ocena algorytmu | Średnia ocena ze scenariusza 1 | Średnia ocena ze scenariusza 2 |
| Puppies\_9000k | 4,92 | 4,15 | 4,08 |
| Puppies\_7000k | 4,79 | 3,62 | 3,77 |
| Puppies\_4000k | 3,96 | 2,92 | 3,00 |
| Puppies\_3000k | 2,33 | 2,15 | 2,31 |
| Puppies\_1500k | 3,04 | 1,62 | 1,46 |
| Chimei\_1500k | 2,63 | 2,62 | 2,69 |
| Chimei\_2000k | 2,08 | 3,23 | 3,15 |
| Puppies\_2000k | 1,00 | 1,77 | 1,54 |

Tabela 6. Zestawienie średnich ocen wystawionych przez algorytm oraz danych z dwóch wybranych algorytmów.

Podczas analizy wyników odrzucono dane pochodzące od jednego z testerów. Powodem była zbyt duża rozbieżność między wystawianymi ocenami, a średnia zbadaną w danej grupie. Obliczone średnie dążą do większych wartości. Spowodowane jest to faktem, iż w sztucznie stworzonym scenariuszu testowym powstałym z wybranych danych otrzymanych podczas badania porównawczego, brakuje sekwencji filmowych o najwyższej jakości. Takie sekwencje wpłynęły na osoby badane w obu scenariuszach ze skalą 5-stopniową.

Otrzymane rezultaty są bardzo mocno zależne od zestawionych par. Stosowanie się do zaleceń przedstawionych na początku rozdziału pozwala na minimalizację błędu metody. Ze względu na zestawienie filmu *Puppies\_1500k* z filmem *Puppies\_7000k*, między którymi zachodziła zbyt duża różnica jakości, średnia ocena algorytmu różni się na tyle, iż postanowiono nie brać go pod uwagę. Inne średnie ocen algorytmu w porównaniu z otrzymanymi bezpośrednio od osób badanych nie różnią się znacząco i mieszczą się w granicach przyjętego błędu. Sekwencje filmowe o gorszych jakościach, zbliżonych do minimalnej użytej w badaniu, oceniane są przez algorytm bardzo precyzyjnie.

Różnice wynikają także z właściwości użytej skali. Skala 7-stopniowa użyta w scenariuszu porównawczym formułuje pytanie testowe w inny sposób. Porównując dwa różne filmy człowiek odpowiada w odmienny sposób, niż w sytuacji, gdy jest pytany o ocenienie jednej sekwencji. Ten fakt powoduje kolejne błędy podczas próby porównania wyników. Należy pamiętać, iż przeprowadzone badania dotyczą kwestii subiektywnych i ich interpretacja jest bardzo trudna.

[POTRZEBA WIĘCEJ WNIOSKÓW]

# Analiza porównawcza