STUDIA INFORMATICA 2010
Volume 31 Number 2B (90)

Danuta ZAKRZEWSKA
Politechnika Łódzka, Instytut Informatyki

ZASTOSOWANIE TECHNIK EKSPLORACJI DANYCH DO BUDOWANIA GRUP STUDENCKICH

Streszczenie. Znalezienie grup studentów o podobnych preferencjach umożliwi dopasowanie do ich potrzeb systemu nauczania na odległość. Celem pracy jest porównanie różnych technik eksploracji danych do budowania grup. Rozważa się zastosowanie klasyfikacji bez nadzoru oraz po nadzorem, jak również wykrywania wzorców sekwencji.

Słowa kluczowe: adaptacyjne systemy nauczania na odległość, grupowanie, klasyfikacja, częste wzorce

DATA MINING TECHNIQUES FOR BUILDING STUDENT GROUPS

Summary. Finding student groups of similar preferences enables to adjust e-learning systems according to their needs. In the paper, it is compared usage of different data mining techniques for creating learners' groups. It is considered application of supervised and unsupervised classification as well as frequent pattern mining.

Keywords: adaptive e-learning systems, grouping, classification, frequent pattern mining

1. Wstęp

Aby osiągnąć oczekiwane efekty kształcenia, system nauczania na odległość powinien być dopasowany do potrzeb studentów. W przypadku dużej liczby użytkowników, dostosowanie oprogramowania do indywidualnych potrzeb może być trudne i kosztowne. Stworzenie grup studentów o podobnych preferencjach pozwala na ograniczenie liczby personalizowanych wersji materiałów dydaktycznych oraz oprogramowania. W pracy bada się możliwość budowania grup studenckich poprzez zastosowanie metod eksploracji danych. Rozważane są trzy techniki: klasyfikacja, analiza klastrowa oraz znajdowanie częstych wzorców. Jako kry-

terium efektywności rozpatrywanych metod przyjęto jakość otrzymanych grup. Badania skupiają się na modelach studentów opartych na stylach uczenia się oraz potrzebach związanych z użytecznością systemu nauczania na odległość.

Pracę zorganizowano w następujący sposób. Sekcja druga została poświęcona personalizacji w systemach nauczania na odległość. W kolejnej części przedstawiono zarówno modele studentów, jak i techniki eksploracji danych wykorzystane do grupowania. Ewaluacja rozpatrywanych metod została przeprowadzona na podstawie eksperymentów opisanych w sekcji czwartej. Na zakończenie zaprezentowano końcowe wnioski oraz plany dotyczące przyszłych badań.

2. Personalizacja w systemach nauczania na odległość

Personalizacja systemów nauczania na odległość stała się w ostatnich latach przedmiotem badań wielu naukowców. Budowanie modeli studentów jest jednym z podstawowych obszarów zainteresowań w tej dziedzinie. Wielu autorów badało cechy charakterystyczne użytkowników, które decydują o ich potrzebach i preferencjach (por. [1] i [2]). Jako jedne z ważniejszych, Brusilovsky [3] wymienił cechy kognitywne, wyznaczane jako wynik testów psychologicznych. Beaudoin [4] zauważył, że dopasowanie środowiska do profili studentów jest warunkiem koniecznym efektywności kursów prowadzonych on-line.

Cechy kognitywne, takie jak style uczenia się, były dość często brane uwagę podczas badań związanych z personalizacją oprogramowania edukacyjnego (por. [5]). Li et al. [6] rozważali dopasowanie do potrzeb studentów zarówno środowiska edukacyjnego, jak i programowego. To ostatnie, ze szczególnym uwzględnieniem dostosowania interfejsu do potrzeb użytkowników, było przedmiotem badań prowadzonych przez Cha et al. [7].

Jednym z czynników, które wpływają na łatwość użycia interfejsu, jest odpowiedni dobór kolorów [8], choć jak dotychczas, niewielu autorów uwzględniało preferencje kolorystyczne podczas budowania modeli użytkowników [9]. Badania przedstawione w pracy [10] pokazały, że możliwe jest stworzenie grup studenckich o podobnych stylach uczenia się i preferencjach kolorystycznych jednocześnie.

Metody eksploracji danych były wykorzystywane do grupowania studentów i budowania personalizowanych systemów nauczania na odległość wielokrotnie. Klasyfikacji używano głównie do zarządzania procesem dydaktycznym [11], predykcji prawidłowości i czasu otrzymania następnej odpowiedzi od studenta [12], a także podejścia studenta do nauki [13], w większości używając danych zawartych w plikach logów.

Analizę klastrową stosowano do grupowania studentów według ich zachowania (por. [14, 15]) czy też stylów kognitywnych ([16]). Tang i McCalla [17] używali analizy klastrowej do

zbudowania systemu rekomendacyjnego, gdzie treść kursu była dopasowana do indywidualnych potrzeb użytkowników. Shen et al. [18], z kolei, zbudowali system rekomendacyjny w oparciu o zachowania użytkowników, wykorzystując zarówno analizę klastrową, jak i technikę odkrywania wzorców sekwencji. Również reguły asocjacyjne były stosowane w systemach rekomendacyjnych na potrzeby edukacji (por. np. [19, 20]). Obszerny przegląd badań dotyczący wykorzystania metod eksploracji danych w systemach edukacyjnych został zaprezentowany w pracy [21].

3. Wykorzystanie metod eksploracji danych

3.1. Modele studentów

Do grupowania zostaną użyte modele studentów oparte na ich stylach uczenia się oraz wymaganiach dotyczących użyteczności, w postaci preferencji kolorystycznych. Spośród modeli opisujących style uczenia się został wybrany model Feldera i Silvermana [22], używany często przy badaniach dotyczących personalizacji systemów nauczania na odległość (por. [23]). Dla każdego studenta model otrzymywany jest na podstawie kwestionariusza ILS (ang. *Index of Learning Styles*) [24], którego wynikiem są preferencje umieszczone w 4 wymiarach, złożonych z 8 wzajemnie wykluczających się par: *aktywny* albo *refleksyjny*, *sensualny* albo *intuicyjny*, *wizualny* albo *werbalny*, *sekwencyjny* albo *globalny*. Wartości atrybutów opisujących każdy styl uczenia mają postać nieparzystych liczb całkowitych z przedziału [−11, 11].

Atrybuty związane z preferencjami kolorystycznymi są typu nominalnego i reprezentują wybór dwóch kompozycji kolorystycznych, spośród wszystkich, które mogą być użyte w interfejsie oprogramowania. Aby nadać wartości tym atrybutom, studenci są zobowiązani dokonać wyboru dwóch preferowanych kolorów i ustawienia ich w ranking.

W dalszej części pracy do budowania grup studentów zostaną użyte 2 modele. W pierwszym, zbudowanym dla celów klasteryzacji, studenci są reprezentowani przez 4 atrybuty typu całkowitego i 2 typu nominalnego:

$$SM = \left(l_{ar}, l_{si}, l_{vv}, l_{sg}, c_1, c_2\right) = (sm_i)_{i=1,\ldots,6} \tag{1}$$

gdzie l_{ar} oznacza punkty dla *aktywnego* (jeśli przyjmuje wartość ujemną) lub *refleksyjnego* (jeśli jest dodatnia) stylu uczenia się i odpowiednio l_{si}, l_{vv}, l_{sg} reprezentują punktację dla pozostałych wymiarów, z wartościami ujemnymi w przypadkach *sensualnych*, *wizualnych* lub *sekwencyjnych* stylów uczenia się, oraz dodatnimi w przypadkach *intuicyjnych*, *werbalnych* lub *globalnych* stylów uczenia się. Atrybuty c_1 i c_2 oznaczają pierwszy i drugi wybór preferowanych kompozycji kolorystycznych i są typu nominalnego.

W drugim modelu, cechy studentów są reprezentowane wyłącznie przez atrybuty o wartościach nominalnych. Dla pierwszych 4 atrybutów otrzymuje się je jako naturalną konsekwencję znaczenia wartości numerycznych modelu (1) (por. [24]). Liczby z przedziałów [5, 11], [−11, −5] oznaczają, że student wykazuje preferencje do jednego z wymiarów, zaś wartości z przedziału [−3, 3] oznaczają, że student jest zbalansowany względem obu wymiarów. Model o wartościach nominalnych przyjmuje postać:

$$SMN = \left(n_{ar}, n_{si}, n_{vv}, n_{sg}, c_1, c_2\right) \tag{2}$$

gdzie

$$n_{ar} = \begin{cases} a & l_{ar} \in [-11,-5] \\ b & l_{ar} \in [-3,3] \\ r & l_{ar} \in [5,11] \end{cases}, \tag{3}$$

$$n_{si} = \begin{cases} s & l_{si} \in [-11,-5] \\ b & l_{sir} \in [-3,3] \\ i & l_{si} \in [5,11] \end{cases}, \tag{4}$$

$$n_{vv} = \begin{cases} vs & l_{vv} \in [-11,-5] \\ b & l_{vv} \in [-3,3] \\ vr & l_{vv} \in [5,11] \end{cases}, \tag{5}$$

$$n_{sg} = \begin{cases} s & l_{sg} \in [-11,-5] \\ b & l_{sg} \in [-3,3] \\ g & l_{sg} \in [5,11] \end{cases}. \tag{6}$$

3.2. Klasyfikacja

Jednym z najczęściej używanych modeli predykcyjnych w edukacji jest Naive Bayes (por. np. [12, 13]). Model ten stanowi najprostszą formę sieci Bayesa [25]. W wyniku algorytmu Naive Bayesa, otrzymuje się rozkład prawdopodobieństwa przydziału do klasy. Metoda ta jest niezwykle efektywna w porównaniu do takich technik klasyfikacji jak sieci neuronowe czy też drzewa decyzyjne [26]. Algorytm opiera się na znanej formule Bayesa:

$$P(G_j \mid SMN) = \frac{P(SMN \mid G_j)P(G_j)}{\sum_{i=1}^{n} P(SMN \mid G_i)P(G_i)}, \qquad j = 1,...,n; \tag{7}$$

gdzie n jest liczbą grup studenckich, SMN przedstawia studentów zgodnie z (2). G_j oznacza zdarzenie przynależności do j-tej grupy, zaś $P(G_j / SMN)$ jest prawdopodobieństwem, że student SMN należy do grupy G_j. Algorytm porównuje cechy nowego studenta, ze wszystkimi znajdującymi się w bazie i oblicza prawdopodobieństwo przynależności do grup. Cechy

decydujące o klasach zależą od nauczycieli, którzy dobierają je stosownie do potrzeb przedmiotu. Mogą być nimi dominujące style uczenia się lub preferencje kolorystyczne.

3.3. Analiza klastrowa

Zagadnienie klasteryzacji studentów reprezentowanych za pomocą (1) zostało opisane w pracy [16]. Przedstawione tam 2 wersje algorytmu pozwalają, podczas grupowania, na nadanie priorytetów poszczególnym atrybutom, a w szczególności na położenie nacisku na preferowane style uczenia się bądź upodobania kolorystyczne.

Wersja podstawowa algorytmu, w której znaczącą rolę odgrywają preferencje dotyczące użyteczności, ma następujące kroki (por. [16]):

Faza I: Klasteryzacja w pojedynczej warstwie:
Dane: atrybuty studentów *SM*, próg klastrowania *T*, maksymalna liczba klastrów *KMAX*, minimalny próg klastrowania *MINT*.
Wyjście: Zbiór klastrów *SCM*.
Kroki:
1. Przydziel SM_1 jako pierwszy klaster, $i=1$;
2. Wyznacz podobieństwo i-tego studenta do środka każdego istniejącego klastra;
3. Przydziel studenta do najbliższego klastra, o ile podobieństwo jest większe niż *T*, wyznacz nowy środek klastra; w przeciwnym wypadku SM_i inicjuje nowy klaster;
4. Powtarzaj kroki 2 i 3 dla każdego studenta SM_i, $i = 2, ...,N$;
5. Jeśli istnieją klastry zawierajace 1 element,powtarzaj kroki 2 i 3, aż klastry się ustabilizują.
Faza II: Technika hierarchiczna aglomeracyjna
Wyjście: Zbiór klastrów *SCM* i zbiór wyjątków.
Kroki:
1. Usuń wszystkie 1-elementowe klastry z *SCM*, niech tworzą zbiór SCM_i;
2. Jeśli liczba klastrów w zbiorze *SCM* jest większa bądź równa *KMAX* idź do 4;
3. Powtarzaj: połącz najbliższe klastry, aż ich liczba osiągnie *KMAX*;
4. Dla każdego klastra z SCM_1, znajdź najbliższy w zbiorze *SCM*, jeśli podobieństwo środków jest większe niż *MINT* połącz je ze sobą, w przeciwnym wypadku wskaż 1-elementowe klastry jako wyjątki

W użytym modelu (1), atrybuty są typu mieszanego: 4 o wartościach numerycznych i 2 o wartościach nominalnych. Dla wyznaczania podobieństwa pomiędzy studentami wykorzystana zostanie funkcja Gowera [27], która umożliwia mierzenie podobieństwa w takim przypadku. Dla dwóch obiektów *i* i *j*, ma ona następującą postać:

$$sim(SM_i, SM_j) = \frac{1}{\sum_{k=1}^{6} w_k} \sum_{k=1}^{6} w_k s_k , \tag{8}$$

gdzie w_k przyjmuje wartości 0 lub 1 w zależności od tego, czy dany atrybut jest używany podczas klasteryzacji. Dla atrybutów o wartościach numerycznych s_k wyznacza się następująco:

$$s_k = 1 - \frac{\left| sm_{i_k} - sm_{j_k} \right|}{R_k}, \; k = 1,...,4; \tag{9}$$

R_k =100 dla k=1,...,4. W przypadku atrybutów o wartościach nominalnych, s_k jest postaci:

$$s_k = \begin{cases} 1 & \textbf{\textit{dla tych samych wartosci}} \\ 0 & \textbf{\textit{w przeciwnym wypadku}} \end{cases}, \; k = 5,6. \tag{10}$$

Rozdzielenie modelu (1) na dwa, w zależności od typu atrybutów: jeden opisujący style uczenia się oraz drugi reprezentujący preferencje kolorystyczne; pozwala na użycie każdej z części modelu w innej fazie. Jeśli początkowe grupy są budowane poprzez użycie preferowanych stylów uczenia się, zaś w fazie drugiej brane są pod uwagę preferencje dotyczące użyteczności, to te ostatnie mają dużo mniejszy wpływ na otrzymany wynik, niż w przypadku zastosowania wszystkich atrybutów w obu fazach.

3.4. Znajdowanie częstych wzorców

Model (2) pozwala na użycie techniki znajdowania częstych wzorców do budowania grup studenckich. Metoda polega na znalezieniu cech studentów, które występują często razem i zbudowania wokół nich grup. Na początek poszukuje się reguł, które łączą wartości atrybutów i występują dostatecznie często w całej populacji studentów. W następnym kroku buduje się grupy poprzez odpowiednie przyporządkowanie studentów do znalezionych wzorców. Jako kryterium przyjmuje się najdłuższy wzorzec charakteryzujący studenta. W przypadku wystąpienia większej liczby wzorców o tej samej długości, przypisanie następuje do grupy reprezentowanej przez regułę o największym wsparciu. Takie podejście pozwala na preferowanie dużych grup studenckich. W przypadku braku możliwości dokładnego dopasowania cech studenta do istniejących wzorców, przydziela się go do grupy według najdłuższej części wspólnej z istniejącymi wzorcami. W efekcie otrzymujemy grupy i ich wzorce. Studenci, dla których nie można znaleźć części wspólnej z istniejącymi wzorcami, muszą być rozpatrywani indywidualnie.

Jednym z najczęściej stosowanych algorytmów do znajdowania częstych wzorców jest „A priori", który opiera się na stwierdzeniu, że dowolny niepusty podzbiór częstych elementów musi być również częsty. W pracy wykorzystany zostanie algorytm typu "A priori", który iteracyjnie redukuje minimalne wsparcie (częstość wystąpienia wzorca), dopóki nie znajdzie wymaganej liczby reguł o zadanej ufności (ufność reguły A $\Rightarrow$ B jest wiarygodnością, że zbiór elementów zawierających A zawiera również B) [28].

4. Eksperymenty i analiza wyników

Celem eksperymentów było porównanie efektywności przydziału studentów do grup, w zależności od zastosowanej metody: klasyfikacji, analizy klastrowej oraz wykorzystania

częstych wzorców, dotyczących atrybutów studentów. Do analiz użyte zostały modele oparte na dominujących stylach uczenia się oraz wymaganiach związanych z użytecznością oprogramowania w postaci preferencji kolorystycznych, wyrażone za pomocą modeli (1) i (2).

Podczas eksperymentów zastosowane zostały 2 zbiory danych: zebrane od 71 rzeczywistych studentów oraz 73 dane wygenerowane sztucznie. Aby otrzymać możliwie jak najbardziej reprezentatywne wyniki testów, specjalny nacisk został położony na zróżnicowanie profili studentów, którzy wzięli udział w badaniach, zakładając, że uczestnicy kursów nauczania na odległość mogą pochodzić z różnych środowisk, dysponować różnym poziomem umiejętności oraz być w różnym wieku. W związku z tym, dane do eksperymentów były zbierane podczas 3 lat akademickich, wśród studentów studiów inżynierskich i magisterskich, stacjonarnych i niestacjonarnych (wieczorowych i zaocznych). Studenci zostali wybrani spośród grupy biorącej udział w międzynarodowej współpracy online przy użyciu środowiska edukacyjnego opartego na *Moodle* (*Open Source*), w ramach projektu CAB programu Socrates Minerva. Wszyscy uczestnicy testów studiowali Informatykę, jednak studenci studiów zaocznych II stopnia posiadali również dyplomy takich kierunków, jak "zarządzanie", "finanse" czy też "marketing".

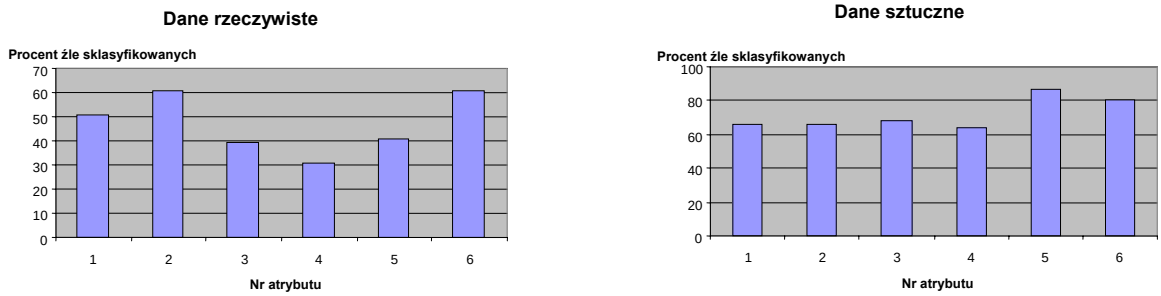
Studenci mieli za zadanie wybrać 2 spośród 9 zaproponowanych im kompozycji kolorystycznych i wskazać najbardziej im odpowiadającą, wypełnili również kwestionariusz ILS. Dane sztuczne zostały wygenerowane losowo, stosując rozkład jednostajny, tak by zachowane zostały typy i zakresy poszczególnych atrybutów.

Za podstawę efektywności rozważanych metod przyjęto błąd przydziału do poszczególnych grup, mierzony liczbą źle sklasyfikowanych studentów.

Do klasyfikacji pod nadzorem wykorzystano algorytm Naive-Bayes, zaimplementowany w programie Weka (*Open Source*) [28]. Klasyfikacja została przeprowadzona dla każdego z 6 atrybutów reprezentujących cechy studentów, obu zbiorów danych, przy wykorzystaniu walidacji krzyżowej. W przypadku pierwszych czterech atrybutów, studenci zostali podzieleni na 3 klasy natomiast dla dwóch ostatnich na 9 klas. Procentowe wartości błędów zostały przedstawione na rys. 1. Dla danych rzeczywistych należą one do przedziału [30.99, 60.56]. Dla danych sztucznych są one wyższe (wszystkie przekraczają 60%). Najmniejszy błąd (30.99%) uzyskano, dla danych rzeczywistych, tworząc klasy dla czwartego wymiaru stylu uczenia się (sekwencyjny/globalny).

Analiza klastrowa studentów opisanych za pomocą modelu (1) została przeprowadzona za pomocą dwóch wersji algorytmu przedstawionego w Sekcji 3.3: podstawowej (wersja 1) oraz rozdzielając atrybuty różnego typu (wersja 2). Również ze względu na mieszany typ atrybutów, analizę jakości otrzymanych grup przeprowadzono w dwóch etapach. W pierwszym zbadany został błąd przydziału do grup, mierzony przez preferencje kolorystyczne, róż-

ne od wskazanych przez większość grupy; w drugim etapie zbadano jakość klastrów z punktu widzenia podobieństwa dominujących stylów uczenia się.



Rys.1. Procent źle sklasyfikowanych przypadków dla danych rzeczywistych i sztucznych
Fig. 1. The percentage of incorrectly classified cases for artificial and real data

Błąd klasyfikacji dotyczący preferencji kolorystycznych został osobno wyznaczony dla pierwszego oraz drugiego wyboru, a także biorąc pod uwagę jedną z dwóch wskazanych kompozycji. Tabela 1 przedstawia procentowe wartości błędów dla rzeczywistego zbioru danych oraz obu wersji algorytmów, w zależności od liczby klastrów. Najmniejszą wartość błędu, dla obu wersji algorytmu, osiągnięto w przypadku 4 klastrów, co oznacza, że tworzenie mniejszych grup nie wpływa na poprawienie ich jakości. Procent źle przydzielonych studentów był niższy niż procent źle sklasyfikowanych przypadków przy użyciu algorytmu Bayesa, choć w dalszym ciągu wysoki: największy dla drugiego wyboru, najmniejszy zaś w przypadku wzięcia pod uwagę 1 koloru, spośród obu wybranych przez studenta. To ostatnie rozwiązanie wydaje się również dobre podczas określania preferencji kolorystycznych reprezentatywnych dla grup.

Tabela 1
Błędy przydziału do grup dla preferencji kolorystycznych - dane rzeczywiste

	Wersja 1			Wersja 2		
Liczba kl.	I wybór	II wybór	I / II wybór	I wybór	II wybór	I/II wybór
3	40.84%	59.15%	25.35%	45.07%	64.79%	28.17%
4	35.21%	54.93%	12.68%	45.05%	64.79%	26.76%
5	32.39%	45.30%	14.08%	45.07%	67.61%	26.76%

Podobnie jak w przypadku klasyfikacji bez nadzoru, wyniki uzyskane dla danych sztucznych były dużo gorszej jakości, Najmniejsza wartość procentowa źle przydzielonych studentów wynosiła 24.7% dla pierwszej wersji algorytmu, przy uwzględnieniu obu wyborów i dla 5 klastrów, maksymalna zaś 67.6% dla drugiej wersji algorytmu, II wyboru i również 5 klastrów. Podobnie jak poprzednio, także dla sztucznych danych liczba źle sklasyfikowanych studentów była mniejsza niż w przypadku klasyfikacji pod nadzorem.

Ze względu na typ liczbowy atrybutów reprezentujących style uczenia się, wysoką jakość klastrów zapewnia właściwy dobór progu klastrowania. Zakres danych opisujących style uczenia się pozwala na dobór progu, tak by zachowane było wymagane podobieństwo pomię-

dzy elementami klastrów. Przykładowo, w celu zapewnienia, by przeciętna różnica pomiędzy wartościami poszczególnych wymiarów stylów uczenia się była mniejsza niż 4, próg podobieństwa nie może być mniejszy niż 0.96. Badania przeprowadzone dla różnych wersji algorytmu pokazały również, że uwzględnienie dodatkowych dwóch atrybutów typu nominalnego pogarszało znacznie otrzymane wyniki.

Ostatnia część testów dotyczyła budowania grup studenckich wokół wzorców reprezentujących atrybuty. Częste wzorce dla obydwu zbiorów danych zostały znalezione przy użyciu algorytmu „A priori" opisanego w punkcie 3.4 i zaimplementowanego w programie Weka [28], przyjmując minimalny poziom ufności równy 0.5. W zbiorze danych rzeczywistych algorytm wykrył 1 wzorzec czteroelementowy, 21 trzyelementowych, 32 dwuelementowych oraz 13 jednoelementowych. W pierwszym etapie zbudowanych zostało 14 grup studentów o tych samych wzorcach, przy czym 3 studentów z rozpatrywanej populacji nie pasowało do żadnej grupy. Ze względu na zbyt dużą liczbę grup, w drugim etapie zostały one ograniczone do 8 przez połączenie tych, których wzorce miały najdłuższą część wspólną. W efekcie powstało 8 grup zawierających od 2 do 23 studentów reprezentowanych przez wzorce o długości od 1 do 3 atrybutów. Największą grupę stanowiło 23 wizualnych studentów, najmniejszą 2 o zbalansowanych stylach uczenia się. Najdłuższy wzorzec dotyczył 7 aktywnych, wizualnych studentów, którzy wybrali kolor nr 2 jako pierwszy preferowany. Kolor ten jako jedyny spośród wszystkich pojawił się we wzorcach czterech grup. W przypadku danych sztucznych algorytm znalazł dużo mniej wzorców, były one krótsze i w żadnym z nich nie wystąpiły preferencje kolorystyczne.

5. Uwagi końcowe

W rozdziale rozważone zostały trzy techniki eksploracji danych, które mogą służyć do klasyfikacji bądź grupowania studentów i których cechy są reprezentowane przez dane nominalne bądź typu mieszanego. Ewaluacja technik została przeprowadzona na podstawie eksperymentów wykonanych na danych rzeczywistych i sztucznych. Wykorzystane zostały modele oparte na stylach uczenia się i preferencjach kolorystycznych, opisanych przez atrybuty typu nominalnego i mieszanego.

Testy pokazały, że dane typu nominalnego generowały większe błędy, mierzone przez procent niewłaściwie sklasyfikowanych przypadków, niż dane typu mieszanego. Klasyfikacja pod nadzorem wykazała się gorszą efektywnością niż klasyfikacja bez nadzoru. Wykorzystanie częstych wzorców do budowania grup powoduje generowanie dużej liczby grup, ich zmniejszenie jest związane z ograniczeniem liczby atrybutów występujących we wzorcach.

Błędy w zakresie atrybutów, które są elementami wzorców, są równe zero, pozostałe atrybuty nie są przy grupowaniu uwzględniane.

Dalsze badania będą dotyczyły rozszerzenia modeli studentów oraz sprawdzenia efektywności innych metod z rozpatrywanych kategorii technik do grupowania studentów w adaptacyjnych systemach nauczania na odległość.

BIBLIOGRAFIA

1. Brusilovsky P., Peylo C.: Adaptive and intelligent web-based educational systems. International Journal of Artificial Intelligence in Education, Vol.13, 2003, s. 156÷169.
2. Stash N., Cristea A., De Bra P.: Authoring of learning styles in adaptive hypermedia: problems and solutions. Proceedings of WWW Conference, New York 2004, s. 114÷123.
3. Brusilovsky P.: Adaptive hypermedia, Use Model. User-Adap., Vol. 11, 2001, s. 87÷110.
4. Beaudoin M. F.: Learning or lurking? Tracking the "invisible" online student. Internet & Higher Education, Vol. 5, 2002, s. 147÷155.
5. Lu J., Yu C. S., Liu C.: Learning style, learning patterns and learning performance in a WebCT-based MIS course. Inform. Manage., Vol. 40, 2003, s. 497÷507.
6. Li Z., Sun Y., Liu M.: A web-based intelligent tutoring system. Artificial Intelligence and Innovations AIAI2005. IFIP International Federation for Information Processing, Vol. 187, Springer, Boston 2005, s. 583÷591.
7. Cha H. J., Kim Y. S., Park S. H., Yoon T. B., Jung Y. M., Lee J.-H.: Learning styles diagnosis based on user interface behaviors for customization of learning interfaces in an intelligent tutoring system. Ikeda M., Ashley K., Chan T.-W., (eds.) ITS2006, LNCS, Vol. 4053, Springer, Heidelberg 2006, s. 513÷524.
8. Marcus A.: Designing graphical interfaces. Unix World, October 1990.
9. Bauersfeld P. F., Slater J. L.: User-oriented color interface design: direct manipulation of color in context. Proceedings of SIGCHI Conference on Human Factors in Computing Systems: Reaching through Technology, New Orleans 2001, s. 417÷418.
10. Zakrzewska D., Wojciechowski A.: Identifying students usability needs in collaborative learning environments. Proceedings of 2008 Conference on Human System Interaction, Kraków 2008, s. 862÷867.
11. Chen G., Liu C., Ou K., Liu B.: Discovering decision knowledge from web log portfolio for managing classsroom processess by applying decision tree and data cube technology. Journal of Educational Computing Research, Vol. 23, 2000, s. 305÷332.
12. Beck J., Woolf B.: High-level student modeling with machine learning. Proceedings of the 5th International Conference on Intelligent Tutoring System, 2000, s. 584÷593.

13. Arroyo I., Woolf B. P.: Inferring learning and attitudes from a Bayesian Network of log file data. Proceedings of the 12th International Conferenec on Artificial Intelligence in Education, 2005, s. 33÷40.
14. Perera D., Kay J., Koprinska I., Yacef K, Zaďane O. R.: Clustering and sequential pattern mining of online collaborative learning data. IEEE T. Knowl. Data En., Vol. 21, 2009, s. 759÷772.
15. Talavera L., Gaudioso E.: Mining student data to characterize similar behavior groups in unstructured collaboration spaces. Workshop on Artificial Intelligence in CSCL. 16th European Conference on Artificial Intelligence, 2004, s. 17÷23.
16. Zakrzewska D.: Cluster analysis in personalized e-learning systems. Nguyen N. T. & Szczerbicki E. (Eds.): Intelligent Systems for Knowledge Management, Studies in Computational Intelligence, Vol. 252, Springer, Berlin Heidelberg 2009, s. 229÷250.
17. Tang T., McCalla G.: Smart recommendation for an evolving e-learning system. International Journal on E-Learning, Vol. 4, 2005, s. 105÷129.
18. Shen R., Han P., Yang F., Yang Q., Huang J.: Data mining and case-based reasoning for distance learning. Journal of Distance Education Technologies, Vol. 1, 2003, s. 46÷58.
19. Minaei-Bidgoli B., Tan P., Punch W.: Mining interesting contrast rules for a web-based educational system. The Twenty-First International Conference on Machine Learning Applications, 2004, s. 1÷8.
20. Wang F.: On using data-mining technology for browsing log file analysis in asynchronous learning environment. Conference on Educational Multimedia, Hypermedia and Telecommunications, 2002, s. 2005÷2006.
21. Romero C., Ventura S.: Educational data mining: a survey from 1995 to 2005. Expert Syst. Appl., Vol. 33, 2007, s. 135÷146.
22. Felder R. M., Silverman L. K.: Learning and teaching styles in engineering education. Eng. Educ., Vol. 78, 1988, s. 674÷681.
23. Viola S. R., Graf S., Kinshuk, Leo T.: Investigating relationships within the index of learning styles: a data driven approach. Interactive Technology & Smart Education, Vol. 4, 2007, s. 7÷18.
24. ILS Questionnaire, http://www.engr.ncsu.edu/learningstyles/ilsweb.html
25. Lowd D., Domingos P.: Naive Bayes models for probability estimation. Proceedings of 22nd International Conference on Machine Learning, Bonn, Germany, 2005.
26. Kotsiantis S. B.: Supervised machine learning: a review of classification. Informatica, Vol. 31, 2007, s. 249÷268.
27. Gower J.: A general coefficient of similarity and some of its properties. Biometrics, Vol. 27, 1971, s. 857÷874.

28. Witten I. H., Frank E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition. Morgan Kaufmann Publishers, San Francisco 2005.

Recenzenci: Dr inż. Paweł Kasprowski
Dr inż. Michał Kawulok

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

Finding student groups of similar preferences enables to adjust e-learning systems according to their needs. In the paper, it was compared the usage of different data mining techniques for finding learners' groups of similar features. There were considered three different data mining techniques: supervised classification by Naive Bayes algorithm (7), clustering technique, which allows to use data of mixed type (see Section 3.3) as well as frequent pattern mining based on the "Appriori" rule. Evaluation of the methods was done for two models based on mixed type attributes (1) and nominal attributes (2) representing student dominant learning styles and usability preferences.

Experiments, done for real and artificial students' data, indicated that data of nominal values generated bigger errors, measured by incorrectly nested instances, than the one of mixed type. Supervised classification showed worse performance than clustering technique (compare Fig. 1 and Table 1). Application of frequent pattern mining allows to avoid errors, but not all the attributes may be included in the patterns, some of the patterns should be shortened to avoid big number of groups created.

Adres

Danuta ZAKRZEWSKA: Politechnika Łódzka, Instytut Informatyki, ul. Wólczańska 215, 90-924 Łódź, Polska, dzakrz@ics.p.lodz.pl .