Nazwa projektu: Narzędzie eksploracji danych

Autorzy specyfikacji:

imię i nazwisko	*nr indeksu*	*grupa*	*adres email*
Kornel Lewandowski	inf94334	czw1145	kornel.lewandowski@gmail.com
Małgorzata Napieraj	inf94347	czw1145	napieraj.malgorzata@gmail.com

Opis projektu

Celem projektu jest napisanie prostego w obsłudze narzędzia eksploracji danych. Narzędzie ma pozwalać na klasyfikację czy ekstrakcję reguł decyzyjnych ze zbioru danych wybranym algorytmem. Dodatkowo, ma umożliwiać wstępne przetwarzanie zbioru wejściowego – dyskretyzację, normalizację, selekcję atrybutów. Ze względu na to, że docelowymi użytkownikami mają być osoby nie specjalizujące się w dziedzinie *data mining*, interfejs użytkownika ma pozwalać na jak najłatwiejszą obsługę, wliczając w to rozbudowaną pomoc kontekstową oraz informowanie o konsekwencjach wykonywanych akcji czy bieżącym stanie procesu eksploracji. System będzie udostępniał kilka wybranych algorytmów eksploracji danych, włączając w to możliwość zmiany parametrów dotyczących danego algorytmu. Ponadto, narzędzie ma umożliwiać import oraz eksport danych w kilku najpopularniejszych formatach plików (XLS, CSV).

Funkcjonalność projektu

Spis wymaganych funkcjonalności z podziałem na obowiązkowe (min. 10) i opcjonalne (min. 5).

Lp	*Opis funkcjonalności*	*Typ*	*Status*
1	Normalizacja i dyskretyzacja danych	obowiązkowa	
2	Selekcja atrybutów	obowiązkowa	
3	Budowa klasyfikatora na podstawie zbioru uczącego (algorytm *NaiveBayes*)	obowiązkowa	
4	Budowa klasyfikatora na podstawie zbioru uczącego (algorytm *J48*)	obowiązkowa	
5	Budowa klasyfikatora na podstawie zbioru uczącego (algorytm *IBk*)	obowiązkowa	
6	Klasyfikacja przykładów na zbudowanym klasyfikatorze	obowiązkowa	
7	Manipulacja parametrami algorytmów klasyfikacji	obowiązkowa	
8	Import danych w formacie XLS/CSV/ARFF	obowiązkowa	
9	Eksport danych w formacie XLS/CSV	obowiązkowa	
10	Pomoc kontekstowa	obowiązkowa	
11	Import/eksport modelu klasyfikatora	opcjonalna	
12	Import danych w formacie XML	opcjonalna	
13	Eksport danych w formacie XML	opcjonalna	
14	Udostępnienie dodatkowego algorytmu selekcji atrybutów (*GainRatio)*	opcjonalna	
15	Budowa klasyfikatora na podstawie zbioru uczącego (algorytm *DTNB*)	opcjonalna	

Ogólny opis poszczególnych grup funkcjonalności

Normalizacja i dyskretyzacja danych – funkcjonalność umożliwi na etapie wczytywania danych pozwoli przeskalować część danych, tak by wszystkie wartości mieściły się w określonym przedziale (pozwoli to łatwo odróżnić bliskie sobie wartości znajdujące się w zbiorze uczącym). Ponadto dane liczbowe będą mogłybyć grupowane w zależnosci od wartości w bardziej czytelne grupy (np. temperatura ciała: niska, w normie, wysoka, bardzo wysoka), co ułatwi zarówno prezentację danych jak i poprawi skuteczność działania klasyfikatorów opartych o przetwarzanie symboliczne.

Selekcja atrybutów pozwoli odrzucić zbędne dla danego procesu klasyfikacji atrybuty decyzyjne, które mogą być zaniżającym jakość klasyfikacji szumem informacyjnym lub danymi nadmiarowymi w kontekście rozpatrywanego przypadku. Opcjonalnie dodane zostaną automatyczne selektory atrybutów, które na podstawie miary entropii warunkowej poszczególnych atrybutów decyzyjnych (ewentualnie również innych miar) będą usuwać atrybuty uznawane za nieprzydatne w procesie klasyfikacji.

Budowa klasyfikatorów oparta będzie o gotowe rozwiązania znajdujące się w pakiecie *Weka*. Każdy z nich sterowany będzie za pomocą parametrów dostępnych w interfejsie użytkownika (oczywiście to, jakie parametry będą ustalane, jest zależne od wybranego algorytmu klasyfikującego). Klasyfikatory zaprojektowane będą w ten sposób, że dane wejściowe (wstępnie przetworzone przez sekecję atrybutów, dyskteryzację wartości i tym podobne) przekonwertowane zostaną do postaci pliku ARFF (domyślny format pliku pakietu *Weka*) i poddane budowaniu modelu klasyfikatora. Na tej podstawie otrzymany zostanie model potrafiący zaklasyfikować nowe przypadki podane przez użytkownika.

Import i eksport danych oparty będzie na prostych konwerterach zrealizowanych za pomocą operacji tekstowych. Konwersja ARFF do CSV (i odwrotna) będzie stosunkowo prosta, ponieważ pliki mają podobną strukturę. W przypadku plików XLS (i być może XML) zostaną wykorzystane zewnętrzne biblioteki.

Pomoc kontekstowa będzie oparta przede wszystkim o komunikaty uświadamiające użytkownikowi konsekwejce jego działań lub proponujące sensowne (z punktu widzenia charakterystyki aktualnie przetwarzanych danych) akcje.

Koncepcja realizacji

Koncepcja realizacji projektu polega na wykorzystaniu istniejącej już biblioteki stworzonej do eksploracji danych oraz stworzenia dla niej wygodnego interfejsu użytkownika. Projekt zakłada wykorzystanie narzędzi eksploracji danych zawartych w pakiecie *Weka*, który udostępnia takie funkcje jak wstępne przetwarzanie zbioru wejściowego, zbudowanie klasyfikatora oraz samą klasyfikację. Pozostałe funkcje, czyli import oraz eksport danych, interfejs użytkownika z pomocą kontekstową i użycie klasyfikatorów zostaną zaimplementowane w ramach projektu.

Rdzeń aplikacji napisany zostanie w języku Java i wykorzystywał będzie biblioteki pakietu *Weka*. Intefrejs użytkownika będzie wykorzystywał przeglądarkę internetową, co pozwoli stworzyć szybko i łatwo atrakcyjne graficznie i funkcjonalne ekrany aplikacji. W celu zapewnienia wysokiej jakości interfejsu wykorzystane zostaną technolowgie związane z aplikacjami przeglądarkowymi: HTML, CSS, JavaScript oraz biblioteka jQuery (ewentualnie jQueryUI).

Połączenie między częścią interfejsu uruchamianą w przeglądarce a rdzeniem aplikacji (programem działającym w tle) zestaw narzędzi Jetty, który to będzie w stanie obsłużyć wysyłanie z przeglądarki żądania i odpowiedzieć na nie komunikatami w odpowienim formacje (JSON).

System w wersji podstawowej będzie obsługiwał 3 różne formaty danych (ARFF, CSV, XLS) oraz udostępniał 3 różne algorytmy eksploracji (*NaiveBayes*, *J48*, *Ibk)* i 1 algorytm selekcji atrybutów *(InfoGain)*.

Materiały źródłowe

- http://www.cs.waikato.ac.nz/ml/weka/documentation.html
- http://api.jquery.com/
- https://wiki.eclipse.org/Jetty/Tutorial/Embedding_Jetty
- http://www.cs.put.poznan.pl/jstefanowski/aed/preprocessing1.pdf
- http://www.cs.put.poznan.pl/jstefanowski/aed/TPDocenaklasyfikatorow.pdf
- http://www.cs.put.poznan.pl/jstefanowski/aed/TPD-grupowanie2011.pdf
- http://www.cs.put.poznan.pl/jstefanowski/ml/W2drzewaJS2010.pdf