Kraków, 31.02.2012

**Projekt z przedmiotu Sztuczne Sieci Neuronowe**

System grupujący wycinki genomów w funkcjonalne grupy

Rafał Babiarz

Joanna Misztal

Przemysław Patelka

1. **Opis zadania:**

Projekt polega na stworzeniu systemu grupującego wycinki neuronów w funkcjonalne grupy za pomocą sieci neuronowych. Jako dane testowe wykorzystane zostały pliki typu .fasta z opisem fragmentów genotypów dla tego samego organizmu. Odcinki genotypu mogą się jednak różnić w ramach tego samego organizmu. Cele zadania jest klasyfikacja genów ze względu na te właśnie różnice.

1. **Opis zastosowanego podejścia:**
2. Wczytywanie i zapis danych z plików do dalszej analizy:

Pliki .fasta wczytywanie są przy pomocy biblioteki "seqRFLP". Następnie każdy genotyp zamieniany jest na swoją binarną reprezentację. Chromosom A zamieniany jest na ciąg 1 0 0 0, T na 0 1 0 0, C na 0 0 1 0, G na 0 0 0 1. Takie podejście powoduje, że żaden chromosom nie jest bardziej podobny do innego (co spełnia wymagania biologiczne projektu). Kolejnym etapem przygotowywania danych jest zastąpienie błędów odczytów (- w pliku fasta) na ciąg 0.25 0.25 0.25 0.25. Zakładamy, że błąd odczytu może oznaczać z takim samym prawdopodobieństem wystąpienie któregokolwiek z chromosomów. Ostatnim etapem jest usunięcie części genotypów, które powtarzają się w każdym z odczytów (to znaczy, jeśli w każdym odczycie na piątej pozycji występuje adenina, to piąta pozycja jest ignorowana przy nauczaniu sieci).

Tak przygotowane dane przekazywane są do nauki przy pomocy sieci Kohonena.

1. Podział danych na klasy:

Do podziału genów na klasy zostały wykorzystane sieci Kohonena.

Mapa samoorganizująca się (sieć Kohonena) jest siecią neuronów, do których przypisane są współrzędne na n - wymiarowej przestrzeni. Nauka sieci polega na zmianie współrzędnych neuronów i dopasowaniu ich do wprowadzonych danych. Sieci takie dostosowują sie zatem do wprowadzonych danych pozwalając tym samym na ich klasyfikacje (podział na grupy), co jest celem naszego zadania.

Podział genów w projekcie następuje etapami - najpierw dane są dzielone za pomocą sieci o wymiarach 2x2 (zaobserwowano, że dane dzielą sie na tyle samo klas niezależnie od wielkości sieci). Następnie sprawdzana jest liczebność genów w powstałych klasach. Jeśli w którejś klasie znajduje sie więcej nic 15% wszystkich genów, to następuje jej kolejny podział siecią Kohonena. Dzieje sie tak aż do momentu, gdy liczebności klas nie przekraczają zadanego progu lub gdy w kolejnym podziale wszystkie geny trafia do tej samej klasy.

W ten sposób można przeanalizować wyniki pod względem kolejnych podziałów. Najprawdopodobniej  początkowo geny zostaną podzielone na bardziej ogólne kategorie, a następnie zostanie dokonany podział ze względu na bardziej szczególne cechy.

1. **Wyniki dla danych testowych:**
2. **Wnioski i obserwacje:**