## Оскільки нейронні мережі не мають людиночитаємого звіту щодо їхної внутрішньої структури, будемо користуватися методом чорного ящика.

## Модель: Використання єдиної нейронної мережі

На роль першого класифікатора було вирішено обрати простий багатошаровий перцептрон Розенблатта завдяки простоті реалізації та наявності багатьох програмних бібліотек, що реалізують необхідний функціонал.

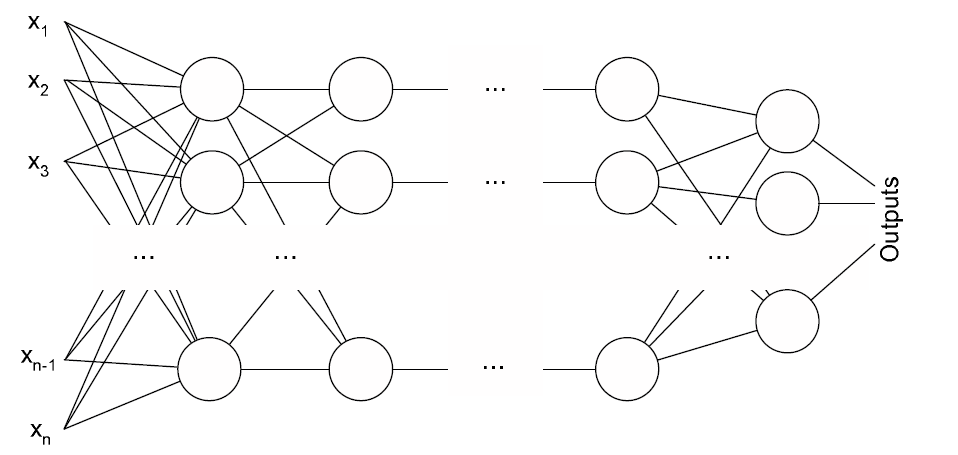


Рис №1 – Багатошаровий перцептрон

Для реалізації було обрано бібліотека FANN (Fast Artificial Neural Network), що спеціалізована на побудові багатошарових нейронних мереж. Оскільки вона написана на мові програмування С, є великий потенціал для портування на різні мови. Ця можливість була використана для біндингу з Java-кодом системи DiaSpectrEx для участі в конкурсі IT-Еврика 2014.

Була обрана повнозв’язкова топологія з кількістю вхідних нейронів, що відповідають кількості ознак в векторі даних та кількістю вихідних нейронів, що відповідають кількості класів. Кількість проміжних шарів та оптимальні функції активації були обрані жадібним алгоритмом на малій частині вихідного набору даних. Класи кодуються таким чином, щоб на виході найбільш ймовірного класу була логічна «1», а на всіх інших виходах – логічний «0».

Для використання перцептрону всі наявні дані були нормалізовані та подані на вхід мережі для її навчання. Також було проведено декілька серій навчання з різними налаштуваннями алгоритмів, та обрано оптимальний з результатів.

## 3.2.3 Тестування моделі

Для тестування обраної технології та її якості навчання на основі набору тренувальних даних було сформовано набір тестових даних за наступним принципом:

https://latex.codecogs.com/gif.latex?v%27_i%20%3D%20v_i%20&plus;%20%5Cdelta_i%2C%20%5Cdelta_i%20%5Cin%20%5B-v_i%5Cdiv25%2C%20v_i%5Cdiv25%5D (9)

де https://latex.codecogs.com/gif.latex?v%27_i - тестовий вектор, https://latex.codecogs.com/gif.latex?v_i - вектор з тренувального набору

Тестування було проведене запуском нейронної мережі на тестових даних з аналізом співвідношення кількості правильно розпізнаних векторів та неправильно розпізнаних векторів

Точність класифікації на тестових векторах складає **від 49.92% до 73.11%** в залежності від тестових даних.

Означений підхід мав декілька суттєвих недоліків, що вимусило відмовитися від подальшої його реалізації. Серед них:

1. Неможливість додавати класи до моделі без перенавчання всієї нейронної мережі;
2. Низька точність класифікації, що неприпустимо в медичних системах;
3. Стабільність моделі залежить від багатьох налаштувань, які нетривіально вибрати оптимально та які залежать від характеристик тренувальної вибірки;
4. Немає можливості детектувати, що клас не належить до жодного з відомих моделі
5. Висока ймовірність перенавчання нейронної мережі через її складність.

## Модель: Використання декількох нейронних мереж

Для того, щоб позбавитися від проблем єдиної нейронної мережі, було вирішено використати декілька менших нейронних мереж у ролі класифікаторів типу «один-проти-всіх».

Кожний з класифікаторів є малим перцептроном, що має 50% зв’язків між нейронами, кількість входів, що відповідає кількості ознак, та лише одному виходу. Таким чином формується N класифікаторів, де N – кількість класів. Кожний класифікатор може відрізнити «свій» клас від усіх інших.

Використовуючи такий підхід з’являється можливість суттєво зменшити кожний з малих класифікаторів, порівняно з одним «великим», особливо у разі великої кількості даних. Класифікатори навчаються за допомогою стандартного інкрементального алгоритму на спеціально адаптованих даних, кожний з яких являє собою файл, в якому дані потрібного класу промарковані символом «1», а дані усіх інших класів – символом «0».

Після побудови класифікаторів вони зорганізуються у глибоке дерево прийняття рішень за допомогою жадібного алгоритму:

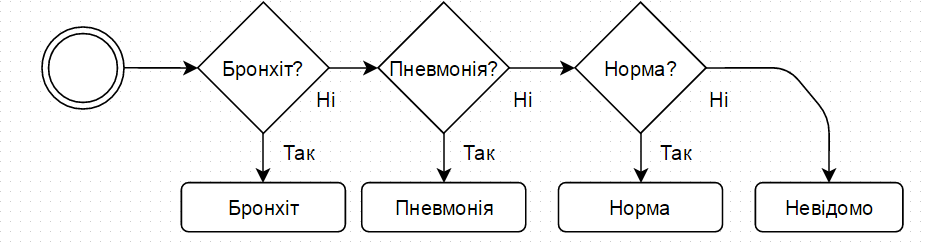


Рис №2 – Примітивне глибоке дерево прийняття рішень

Оскільки кожний з підкласифікаторів представляє собою окрему нейронну мережу, для нього можуть бути використані індивідуальні налаштування, що дозволять виокремлювати «свій» клас найбільш ефективно.

Алгоритм побудови являє собою повний перебір можливих формацій дерева з подальшим тестуванням отриманого дерева на кожній ітерації перестановок блоків-класифікаторів. Використана стратегія «розподіляй і володарюй» допомогла в суттєвому зменшенні складності кожного з класифікаторів та всієї моделі в цілому.

## 3.4 Тестування моделі

Модель добре проявила свої властивості на тестових даних, що були згенеровані так само, як і дані для тестування єдиної мережі (9). Точність розпізнавання складає **94.83%** на тестових даних. Для сучасних медичних досліджень це є досить високим результатом, оскільки ставить технологію на один рівень точності з такими неінвазійними методами діагностики як рентген та флюороскопія.

Однак, не зважаючи на високу точність алгоритму, він має декілька суттєвих недоліків.

З боку навчання:

* Висока обчислювальна складність повного перебору, що унеможливить побудову моделі для великої кількості класів, або для великої кількості даних в класах.
* Неможливість коригування окремих підкласифікаторів без повторного навчання усіх підкласифікаторів моделі.

З боку виконання:

* Неможливість додавання нових класів при роботі алгоритму.
* Нестабільність моделі: побудова моделі залежить від якості навчання кожного з обраних класифікаторів.

Оскільки на даний момент кількість класів досить мала через невелику кількість клінічних досліджень, алгоритм, що використовує дерево рішень є оптимальним с точки зору співвідношення «витрати – якість розпізнавання».

Тестування продукту вцілому:

|  |  |  |
| --- | --- | --- |
| Входные события | Допустимые классы эквивалентности | Недопустимые классы эквивалентности |
| Ввод ФИО | Строковые значения, состоящие из украинских и латинских букв, дефиса и апострофа, не более 60 символов в длину.  Тестовое значение:  Вильгельмина Д’Абдурахмангаджи Христорождественская-Петренко | Числа, специальные символы, типографские символы  Тестовое значение:  Кон©тантин 1 Радомирович Тесленко~Павлов |
| Ввод КВВВ-данных | Дробные значения в диапазоне от 0 до 100, не более 4 знаков после запятой. Десятичный разделитель – запятая  Тестовое значение:  99,3544 | Строки, специальные символы  Тестовое значение:  Девяносто девять,3544 |
| Ввод e-mail | Строка, соответствующая шаблону написания адреса электронной почты  Тестовое значение:  [tim4bor@gmail.com](mailto:tim4bor@gmail.com) | Специальные символы  [tim~4bor@gmail.com](mailto:Tim~4bor@gmail.com)  [tim4bor@.com](mailto:Tim4bor@.com) |
| Ввод номера телефона | Строка, состоящая из цифр и знака «+», длиной 13 символов  Тестовое значение:  +380664471908 | Буквенно-цифровые строки, специальные символы  Тестовое значение:  +38066ЧЧ71908  +3966445585  +369535465© |
| Ввод номера паспорта | Буквенно-цифровая строка, соответствующая шаблону написания номера паспорта  Тестовое значение:  АА123456 | Специальные символы  Тестовое значение:  С©Ё!»№;% |
| Ввод истории болезни | Текст  Тестовое значение:  Lorem ipsum dolor sit amet, consectetur adipiscing elit. Cras imperdiet volutpat massa vel bibendum. Etiam viverra tristique lorem vel tempor. Donec magna neque, condimentum sit amet tincidunt venenatis, faucibus in risus. Nam felis dui, dapibus non viverra vel, lacinia commodo velit. Nam ac sollicitudin nisi. Class aptent taciti sociosqu ad. | Нетипографские символы |
| Ввод даты | Дата в формате ДД-ММ-ГГГГ  Тестовое значение:  23-12-1925 | Буквенные символы, специальные символы,  дата, больше текущей, дата раньше текущей более, чем на 150 лет  Тестовое значение:  АЧ~44-9252  23-02-1800  23-02-2016 |