computing@tanet.edu.te.ua
www.tanet.edu.te.ua/computing

ISSN 1727-6209
International Scientific
Journal of Computing

АНАЛІЗ СТРУКТУРИ МЕДИЧНИХ ДАНИХ ІЗ ЗАСТОСУВАННЯМ МЕРЕЖ КОХОНЕНА

О. В. Годич [2], Ю. В. Нікольський [1], В. В. Пасічник [1], Ю. М. Щербина [2]

[1] Національний університет „Львівська політехніка"
[2] Національний університет ім.І.Франка

Анотація: *У цій статті авторами обговорено підходи до виявлення структури даних високої розмірності із використанням нейромереж, що самоорганізуються. Висвітлені підходи використовують графічні зображення для інтерпретації структури даних. Апробація успішності застосування запропонованих технологій аналізу структур даних проведена на медичних даних, які стосуються пацієнтів із кардіологічними захворюваннями. Аналіз структури медичних даних є продовженням попередніх досліджень авторів з метою розробки ефективної технології діагностування у медицині.*

Ключові слова: *діагностування, кластеризація, класифікація, штучні нейронні мережі, візуалізація даних.*

ПОСТАНОВКА ЗАДАЧІ В ЗАГАЛЬНОМУ ВИГЛЯДІ

Для аналізу медичних даних, орієнтованого на пошук закономірностей утворення груп пацієнтів, можна використати технології машинного навчання, пов'язані із групуванням понять [1], зокрема, методи кластеризації даних. Застосування цих методів дає змогу визначати структури даних їхнім поділом на групи та подальшим виділенням визначальних ознак, за якими можна розрізняти такі групи. Ці ознаки виділяють з множин ознак, якими є симптоми хвороб або результати аналізів, отриманих на етапі обстеження хворого, що передує встановленню діагнозу та призначенню лікування. Метою проведеного дослідження є кластеризація медичних даних та виявлення на цій основі прихованих в них закономірностей.

Проаналізовані дані формують *таблицю прийняття рішень* $B = (Z, A \cup \{d\})$ [2], де Z – множина всіх об'єктів досліджень – пацієнтів, а $A = (a_1, a_2, \ldots, a_m)$ – кортеж [3] атрибутів (симптомів та результатів обстежень), які встановлюють відповідність вигляду $a : Z \to V_a$, де V_a – домен атрибуту $a \in A$, d – *атрибут прийняття рішень*. Атрибути множини A називають *умовними*, або умовами, а d – *рішенням*. Таблицю $B = (Z, A)$, отриману з таблиці прийняття рішень вилученням атрибуту прийняття рішень, називають *інформаційною системою*. Кожний рядок $z = (z_1, z_2, \ldots, z_m)$, $z \in Z$ інформаційної системи називатимемо *прикладом*. Приклад є кортежем, елементами якого є значення відповідних атрибутів кортежу A.

У нашому випадку домени атрибутів є такими, що $V_{a_i} = \{0; 1\}$ $(i = \overline{1, m})$, $V_d = \{0; 1\}$. Зауважимо, що для кластеризації використовують дані з інформаційної системи, тобто приклади з $B = (Z, A)$.

Кластеризація полягає у побудові груп об'єктів у певному сенсі подібних між собою. Групування виконують на основі значень спеціальної функції, яка є мірою подібності об'єктів. Отримані групи називають *кластерами*.

Існують різні підходи до вирішення задачі кластеризації [4]. На рис.1 [6] зображено процес кластеризації множини об'єктів $\{a, b, c, d, e\}$, де дендрограма візуалізує процеси об'єднання та поділу кластерів, що є результатом ієрархічної кластеризації.

Нейромережі типу SOM (*self-organising maps*) є ще однією ефективною технологією виявлення кластерів у даних та дослідження структур даних у задачах видобування даних (data mining and knowledge discovery [7]). Ці нейромережі особливо привабливі для аналізу даних завдяки їхній здатності візуалізації результатів аналізу. Алгоритми навчання нейромереж SOM встановлюють відповідність прикладам високої

розмірності нейрони розташовані у вузлах ґратки меншої розмірності. Кожному нейрону поставлено у відповідність вектор дійсних чисел, який називають ваговим.

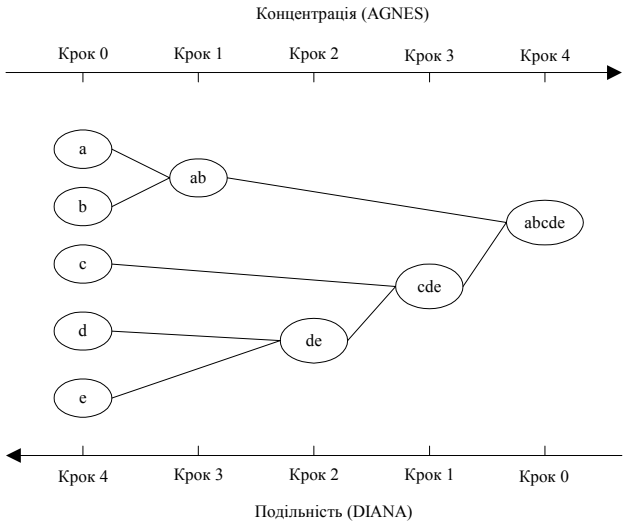

Рис. 1 – Ілюстрація підходів до реалізації процесу кластеризації, основаних на ієрархічних подільності та концентрації

Впорядковані за певним принципом нейрони ґратки використовують для візуалізації властивостей даних, зокрема, їхньої кластерної структури. Таку візуалізацію даних можна використати для визначення якісних характеристик об'єктів (*qualitative information*).

Нейромережа є моделлю, яка відображає структуру множини прикладів інформаційної системи. Кластерна структура множини прикладів інформаційної системи визначає розбиття множини вагових векторів нейронів. Структурні закономірності розподілу прикладів інформаційної системи апріорі невідомі, а їх визначення є проблемою, яку вирішуємо навчанням мережі SOM. Процес навчання використовує тільки ці закономірності без додаткових припущень або інформації про задачу, що вирішується.

Мережа SOM, яку було використано для аналізу медичних даних, складається з m вхідних нейронів та ґратки нейронів. Приклад, який подають на вхід нейромережі є *вхідним вектором*. Кластери з нейронів мережі утворюються впродовж навчання застосуванням відомих алгоритмів послідовного уточнення вагових векторів. Обчислення організовано так, що кластерам інформаційної системи відповідатимуть кластери вагових векторів нейронів у ґратці. Кожному прикладу та близьким йому прикладам, які утворюють кластери, відповідає група нейронів з близькими ваговими векторами. Упродовж навчання кожний вхідний вектор порівнюють із ваговими векторами усіх нейронів, серед яких вибирають той, що задовольняє умову екстремуму певної цільової функції. Нейрон, для якого виконана ця умова, називають *нейроном-переможцем*. Очевидно, що нейрон-переможець існує для довільного вхідного вектора. Як цільову функцію використовують скалярний добуток вхідного та вагових векторів або евклідову відстань між ними. Нейрон-переможець визначає своє *топологічне сусідство* як множину нейронів, близьких до нього за введеною відстанню. Нейрони з топологічного сусідства навчають за вибраним алгоритмом. Мережу вважають навченою, якщо виконані певні умови закінчення алгоритму навчання. У подальшому викладенні, нейромережу SOM також називатимемо мережею Кохонена, оскільки її навчання здійснене за алгоритмом Кохонена [8].

АНАЛІЗ ОСТАННІХ ДОСЛІДЖЕНЬ

Основний напрямок досліджень із використанням SOM пов'язаний із аналізом неструктурованих та слабко структуррованих даних, виявленням прихованих структур та візуалізації процедур та результатів аналізу. В останні роки значна кількість праць із використанням нейромереж SOM присвячена аналізу текстової інформації. Серед таких досліджень цікаво відмітити роботи з аналізу документів, виконаних природними мовами. Традиційний обчислювальний аналіз тексту на природній мові часто зосереджений на її синтаксичній структурі. Одночасно ігнорують загальний контекст тексту, який є об'єктом аналізу. Головна увага у дослідженнях з використанням SOM приділяється не аналізу мовних конструкцій текстів, а взаємозалежності різних текстових документів за контекстом. Аналізом великих обсягів документів робляться спроби знаходження концептуальних, внутрішньо-системних залежностей, які можуть бути основою змістовного навантаження слів.

Автори праці [9] запропонували підхід до аналізу текстових документів із різних експертних галузей, що дає можливість знаходити тематично схожі тексти без жодної попередньої інформації про них. Отримані результати вказують на те, що достатньо простий аналіз ключових слів у комбінації із нейромережею SOM може бути дуже ефективним для визначення контекстної залежності між текстами та їх авторами.

У наукових колах активно обговорюють поширення та роль міждисциплінарних досліджень. Нажаль, не існує хорошого засобу

для вимірювання міри поширення таких досліджень та їхньої важливості для розвитку науки. У праці [10] обговорено цю проблему, а авторами запропоновано новий апарат вимірювання міждисциплінарних досліджень, який базується на нейромережі SOM. Для аналізу вмісту наукових публікацій та документів автори використовували метод WEBSOM.

Нейромережі SOM зарекомендували себе як гнучкий апарат, ефективно застосовний для різних прикладних проблем. Зокрема, у праці [11] авторами досліджена можливість застосування нейромереж SOM для визначення та опису онтологічних зв'язків між семантичними об'єктами та класами об'єктів у базі візуальних об'єктів. Вивчались так онтологічні зв'язки як співіснування, таксономія візуальної та семантичної подібності, а також просторові зв'язки. У цьому досліджені авторами використано базу даних із 2618 зображень.

У праці [12] досліджено можливість використання SOM для системи прийняття рішень DERSI, яку призначено для автоматизації прийняття рішень для визначення дефектів. Кінцевою метою цього дослідження є розробка комп'ютеризованої системи прийняття рішень для атомної електростанції. Однією із особливостей системи DERSI є наявність людино-машинної інтерфейсної частини, яку ефективно використовують для прийняття рішень.

Низка досліджень щодо можливості використання штучних нейронних мереж, зокрема SOM, у медицині висвітлено у [13]. У цій праці проаналізовано успішність декількох застосувань штучних нейронних мереж для діагностування хворих на ракові захворювання.

ЦІЛІ СТАТТІ

Метою статті є висвітлення ефективних способів візуалізації результатів застосування навченої нейромережі Кохонена для аналізу структури даних. Така задача вирішувалась у контексті дослідження моделей прийняття рішень у медицині з метою прогнозування діагнозу захворювань в імунології та кардіології. У результаті проведених досліджень вдається виявити певні закономірності встановлення діагнозу шляхом класифікації пацієнтів на основі аналізу груп симптомів захворювань та результатів обстежень. Дослідження здійснювались застосуванням нейромереж Кохонена для групування даних як без урахування атрибуту прийняття рішень, так й із його урахуванням.

У статті проведено порівняння різних способів візуалізації результатів кластеризації багатовимірних даних, виконаного за допомогою нейромереж, що самоорганізуються. Такий підхід дає змогу виявляти можливі відхилення у встановленні діагнозу, вдосконалювати цей процес та коректувати призначення лікування.

ОСНОВНИЙ МАТЕРІАЛ

Авторами вже проводились дослідження з метою побудови класифікатора для встановлення діагнозу із використанням нейромережі SOM, навченої різними алгоритмами на прикладах інформаційної системи $B = (Z, A)$ з даними, на основі яких встановлено діагноз певного кардіозахворювання. Пошук закономірностей виконано на даних, зібраних впродовж обстеження пацієнтів з метою виявлення передумов виникнення певного кардіологічного захворювання.

Для аналізу використано дані про 3532 пацієнти, зібрані у таблицю з 13 атрибутами, кожен з яких відповідає певному симптому. Діагнозом є значення атрибуту прийняття рішень, позначене одиницею або нулем, що відповідає наявності або відсутності захворювання, відповідно.

Кожний атрибут має значення 0 або 1, стать пацієнтів позначена одиницею для жінок, та нулем – для чоловіків. За віком пацієнти розбиті на дві групи: пацієнти до 50 років позначаються у базі одиницею, а молодші – нулем. Ці дані автори вже аналізували із застосуванням інших методів [14-15].

Навчання нейромережі Кохонена детально досліджено, а його результати опубліковано у праці [8]. У продовження цих досліджень наведемо можливі підходи до візуалізації процедур прийняття рішень, які виконуються з допомогою навченої мережі Кохонена.

Для кластеризації вхідних даних навчено нейромережу Кохонена, проведено детальне дослідження альтернативних алгоритмів навчання такої мережі та знайдено оптимальні значення параметрів цих алгоритмів для досліджуваних даних. Запропоновано алгоритм побудови класифікатора, який використано для автоматичного встановлення діагнозу [8].

Для навчання нейромережі та оцінки якості роботи побудованого на її основі класифікатора, вся множина даних про пацієнтів розбита на дві частини – навчальну χ та перевірочну $Z \setminus \chi$

Навчена нейромережа Кохонена дала змогу встановити певні закономірності у множині прикладів інформаційної системи. Використанням відомих значень атрибуту прийняття рішень відповідної системи прийняття рішень встановлено підмножини прикладів, яким відповідає кожний

нейрон нейромережі Кохонена. Надання нейронам номера кластера відповідних йому прикладів називають *маркуВанням*, а нейромережу, кожний нейрон якої отримав номер кластера, назвемо *промаркованою.*

З допомогою промаркованої мережі Кохонена побудовано класифікатор, який дозволив встановити кожному прикладу з множини Z значення його атрибуту прийняття рішень, а, отже, діагноз відповідного пацієнта.

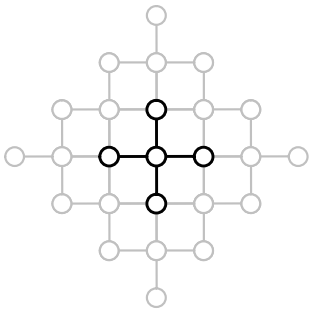

Рис. 2 – Фрагмент нейромережі Кохонена із хрестовидним патерном сусідством

Аналіз даних орієнтований на пошук „природних" кластерів, для кожного з яких характерна певна група симптомів. Оцінена кількість кластерів у множині прикладів інформаційної системи. Використання нейромережі Кохонена дає змогу здійснити таке оцінювання візуально на плоскому зображенні ґратки нейронів. Візуалізацію результатів кластеризація навченою нейромережею Кохонена виконано алгоритмом, який називається *алгоритмом U-Matrix* [16].

Дослідження навченої нейромережі Кохонена виконано у такій послідовності.

1. Обчислення висот нейронів та побудова карти висот (U-Matrix).
2. Проведення експериментів з навченою мережею на навчальній та тестовій множині прикладів та побудова карти частот.
3. Маркування нейронів на тестових та навчальних прикладах, побудова промаркованих ґраток нейронів та побудови відповідних карт.
4. Візуалізація карти частот.

Висотою нейрона називають середнє арифметичне суми відстаней ваг від певного нейрона до нейронів з його безпосереднього сусідства [8] та обчислюють за формулою

$$U_h(i) = \frac{1}{|N(i)|} \sum_{j \in N(i)} d(w_i, w_j),$$

де i – номер нейрона в ґратці, для якого обчислюють висоту, $N(i)$ – множина нейронів із безпосереднього сусідства нейрона з номером i, $|N(i)|$ – потужність множини $N(i)$, $d(\cdot,\cdot)$ – відстань між нейронами з ваговими векторами u та v, яку використано в алгоритмі Кохонена для навчання нейромережі.

Безпосереднє сусідство нейрона залежить від розташування нейронів у ґратці, яке ще називають *патерном сусідства*. Одним із патернів сусідства є хрестовидний. Фрагмент ґратки для мережі Кохонена із хрестовидним патерном сусідства зображено на рис. 2. Такий патерн сусідства використаний у дослідженнях, результати яких наведено у цій статті. Що більше значення висоти нейрона, то більша відмінність між даними, за відображення яких відповідає цей нейрон та нейрони з його безпосереднього сусідства. Відносно велика відмінність значень висоти двох нейронів свідчить про їхню потенційну належність до різних кластерів. Якщо ж ця відмінність є малою, то такі дані потенційно належать одному кластеру.

За обчисленими висотами нейронів навченої нейромережі Кохонена побудоване її візуальне зображення, яке називатимемо *картою висот*. Карта висот накладена на вибраний патерн сусідства ґратки нейронів. Висота нейрона зображена на карті висот відтінком сірого, інтенсивність якого є вищою для більшого значення висоти. Карту висот можна розмежувати на „хребти" та „долини", де „долини" відповідають малим значенням висот нейронів, а „хребти" – великим значенням.

Візуалізація значень опрацьованої за алгоритмом U-Matrix навченої нейромережі Кохонена дає змогу оцінити кількість „природних" кластерів у вхідних даних.

Дослідження закономірностей, які містять дані з використанням навчання нейромережі проведено у такій послідовності.

1. Навчено нейромережі за алгоритмом Кохонена з параметрами, значення яких вибрано за результатами, наведеними у статті [8].
2. Розроблено методику та проведено візуалізацію результатів експериментів за алгоритмом U-Matrix.
3. Оцінено якість навченої мережі за значеннями спеціальних параметрів:

- *середньоквадратичне відхилення* (*MSE, mean square error*);
- *топографічну помилку* (*TE, topographic error*);
- успішність розпізнавання, як відсоток вхідних прикладів, правильно класифікований нейромережею (*PSR, percentage of successful recognitions*).

У циклі досліджень [8], присвячених, зокрема, й побудові класифікатора, використана нейромережа Кохонена, яка складалась зі 100 нейронів з прямокутним патерном сусідства. В результаті досліджень підібрано значення параметрів алгоритму навчання. Застосуванням класифікатора, побудованого на основі навченої нейромережі Кохонена було досягнуто до 93% успішно класифікованих прикладів. Отримання такого результату класифікації стало підставою для застосування для такої мережі алгоритму U-Matrix з метою побудови карти висот. Головна відмінність проведених експериментів, результати яких наведено у цій статтіє зміна патерна сусідства з прямокутного на хрестовидний.

Відповідно до алгоритму побудови класифікатора [8] здійснено *маркування* навченої нейромережі та отримано промарковану нейромережу, яку можна використати як класифікатор. Кожен приклад віднесений до одного з двох класів. Маркування нейронів у ґратці дало змогу виконати візуалізацію розподілу нейронів за значенням атрибуту прийняття рішень. Така візуалізація виконана для двох наборів даних – навчального та тестового. Чорним кольором на маркованих картах замальовано нейрони, які відповідають хворим пацієнтам.

Казатимемо, що нейрон навченої нейромережі Кохонена реагує на вхідний вектор, якщо він є нейроном-переможцем для цього вектора. Критерієм надання мітки нейронам є кількість реагувань нейрона на дані певного класу. Якщо нейрон частіше реагував на дані, що відповідають хворим пацієнтам, ніж здоровим, то такий нейрон маркуємо одиницею.

Проведена низка експериментів із використанням чотирьох нейромереж Кохонена, які складались із 113, 545, 1201 та 2665 нейронів. Для зручності, називатимемо нейромережі за кількістю нейронів. Поступове збільшення кількості нейронів дало змогу прослідкувати закономірності зміни властивостей навчених нейромереж. Зауважимо, що приклади побудованих карт висот, частот та промарковані карти, наведено лише для найбільшої нейромережі, яка містила 2665 нейронів.

Карту частот будують за кількістю реагувань нейронів-переможців на даних із навчальної та тестової множин. Діаметр круга, яким позначено відповідний нейрон у ґратці, пропорційний кількості реагувань, а мітка – число, яке стоїть поруч і вказує точну кількість реагувань. Таке подання дає змогу спостерігати найактивніші нейрони у ґратці, які здебільшого є центрами кластерів.

Для нейромережі 113 за картою висот не можна оцінити кількості кластерів; переважна кількість реагувань припадає на невелику кількість нейронів, яку можна оцінити в 10-15 нейронів; у результаті маркування нейронів, яке проводилось із використанням атрибуту прийняття рішень, отримані зображення практично співпадають для навчальних та тестових прикладів. Також звернемо увагу на те, що нейрони, промарковані як такі, що розпізнають хворих, зосередились у центральній частині ґратки із незначним розсіянням від її центру.

Збільшення кількості нейронів у нейромережі Кохонена покращує візуальне сприйняття інформації про кластерну структуру даних. Для нейромережі 545, карта висот дає змогу більш чітко, ніж на карті висот нейромережі 113, побачити області, які відповідні різним кластерам.

На карті частот нейромережі 545 можна спостерігати приблизно ту саму кількість найактивніших нейронів, що й у нейромережі 113. Кількість таких нейронів також знаходиться в межах від 10 до 15, причому кількість реагувань найактивніших нейронів співпадає в обох нейромереж.

Карти, отримані маркуванням на навчальних та тестових прикладах нейронів мережі 545 мають конфігурації, схожі із відповідними конфігураціями маркованих областей нейромережі 113. Відмінною рисою промаркованих карт нейромережі 545 є збільшення кількості нейронів, які відхиляються від її центральної частини.

Наступною досліджено нейромережу 1201. На карті висот можна візуально розмежувати кластери. Кількість реагувань нейронів на вхідні вектори, що спостерігаються за картою частот у порівнянні з попередніми нейромережами збереглась. Незважаючи на збільшення загальної кількості нейронів, кількість найактивніших нейронів близька до кількості найактивніших нейронів нейромереж 113 та 545.

Конфігурація промаркованих нейронів на картах свідчить про посилення тенденції до розсіювання від центру нейромережевої ґратки нейронів, відповідальних за розпізнавання хворих.

Результати аналізу структури інформації для навченої нейромережі 2665 подані на рис. 3-6. Ця мережа виявила у навчальних та тестових даних такі закономірності:

1. Карта висот на рис. 3 є найчіткішою у порівняння із тими картами, що були отримані для нейромереж із меншою кількістю нейронів.

2. Кількість нейронів із високою активністю (див. рис. 4) близька до тієї, що й у нейромережах з меншою кількістю нейронів.
3. Частоти реагувань у найактивніших нейронів практично співпадають.
4. На промаркованих картах як на навчальній, так і на тестових множинах (див. рис. 5, 6) розсіяння нейронів від центру ґратки збільшилось.

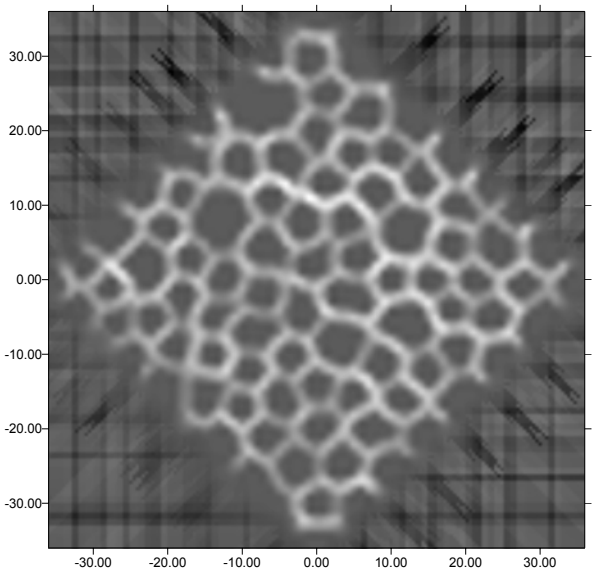

Рис. 3 – Карта висот нейромережі 2665

На карті висот з рис. 3 спостерігається велика кількість природних кластерів. Ґрунтуючись на досвіді лікарів-кардіологів можна обґрунтувати належність пацієнтів до конкретних кластерів та використати отримані результати для оцінювання рівня суб'єктивності встановлення діагнозу.

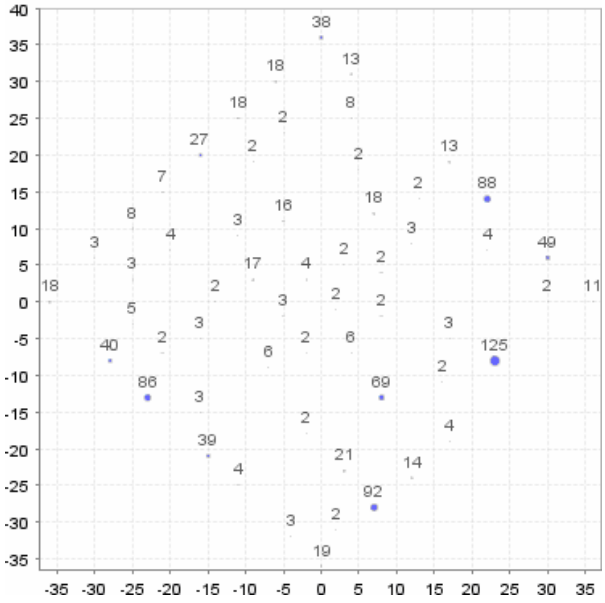

Рис. 4 – Карта частот нейромережі 2665, отримана для навчальної множини прикладів

Результати, отримані з використанням навчених нейромереж Кохонена з різною кількістю нейронів, вимагають додаткового аналізу та інтерпретації.

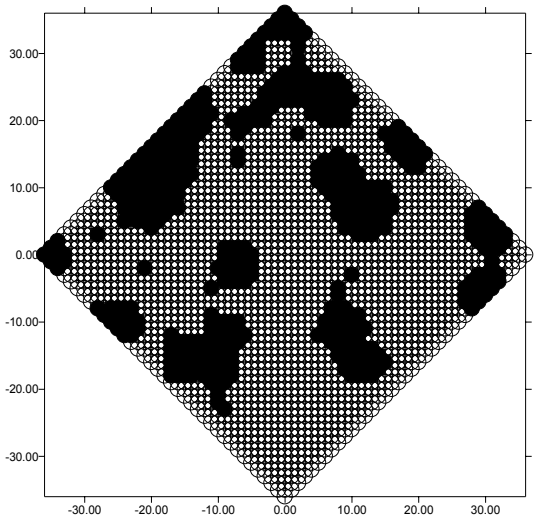

Рис. 5 – Нейромережа 2665, промаркована на навчальній множині прикладів

Це викликано тим, що збільшення кількості нейронів у навченій нейромережі призводить, з одного боку, до значного покращення карти висот, а з іншого – до погіршення промаркованих карт.

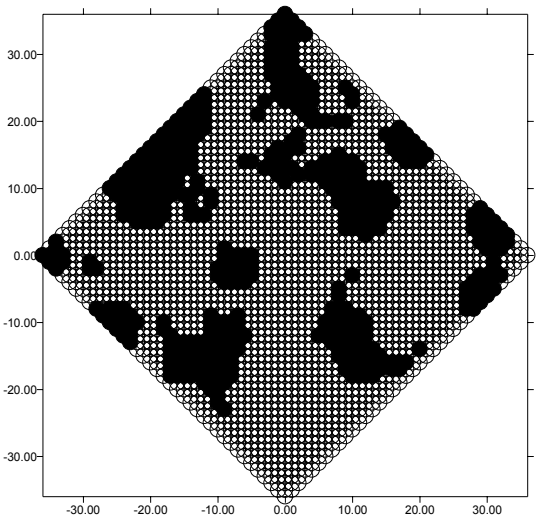

Рис. 6 – Нейромережа 2665, промаркована на тестовій множині прикладів.

У ідеальному випадку нейрони, що відповідають за класифікацію пацієнтів з одного класу, повинні бути близькими у ґратці. Це є результатом однієї із головних властивостей нейромереж SOM – їх здатності до топологічного впорядкування впродовж навчання. Слід зазначити, що формування карти висот відбувається у природний спосіб, тобто, не виконано додаткового опрацювання прикладів інформаційної системи та інтерпретації даних.

Водночас процес маркування здійснено із врахуванням значень атрибуту прийняття рішень, який несе значну міру суб'єктивності. З огляду на це, виконаємо додатковий аналіз навчених нейромереж із використанням двох, незалежних від атрибуту прийняття рішення, параметрів MSE та TE [17], які є традиційними для нейромереж Кохонена числовими характеристиками якості навчання.

Середньоквадратичне відхилення MSE дає змогу оцінити якість апроксимації вхідних даних навченою нейромережею. Цією оцінкою є величина відхилення вхідного вектора $x \in \chi$ від вагового вектора відповідного йому нейрона-переможця.

Отже, за значенням параметра MSE оцінена якість апроксимації множини векторів $\chi \subset Z$ нейромережею. Параметр MSE обчислюють за формулою

$$MSE = \frac{1}{N}\sum_{z\in\chi}\left\|z - w_{BMU}\right\|^2,$$

де w_{BMU} – ваговий вектор нейрона-переможця, або *нейрона найкращого наближення* (*BMU, best matching unit*) для вхідного вектора $z \in \chi$, N – кількість векторів у множині χ. Менше значення MSE вказує на кращу результативність, тобто більш точне моделювання вхідної множини прикладів навченою SOM у сенсі квадратичного відхилення вхідних та вагових векторів.

Значення параметру MSE для всіх нейромереж, результати досліджень яких наведено у цій статті, обчислені для тестових прикладів (див. рис. 7). З цього рисунка видно, що якість апроксимації вхідних даних покращується із збільшенням кількості нейронів у нейромережі. Цей висновок справджується для тестових і навчальних прикладів.

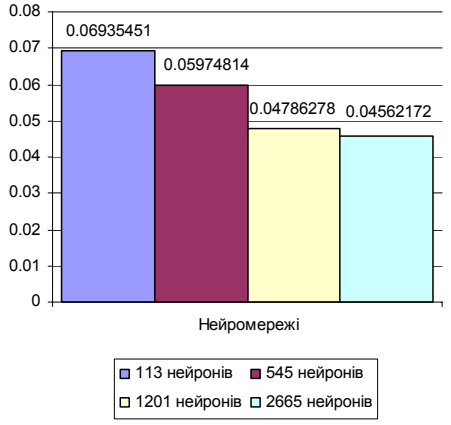

Рис. 7 – Залежність значення параметра MSE від розмірів гратки на тестових прикладах

Другим параметром, використаним для оцінювання якості навчання, є топографічна помилка TE. За значенням TE оцінюють якість впорядкування нейронів у ґратці, тобто невідповідність між розташуванням нейронів нейромережевої ґратки та особливостями вхідних даних. Менше значення цього параметру свідчить про кращу впорядкованість нейронів навченої нейромережі.

Зміст цього параметра треба розуміти як здатність нейромережі Кохонена зображати близькі вектори близькими нейронами. Зображення близьких векторів у нейромережі Кохонена далекими нейронами називають *топологічним збуренням*. Збільшення топологічних збурень зменшує ефективність мережі Кохонена адекватно відображати дані високої розмірності двовимірною картою.

Значення параметра TE обчислюють різними способами. Тут використана формула

$$TE = \frac{1}{N}\sum_{z\in\chi}\begin{cases}1, \left\|r_{BMU} - r_{SBMU}\right\| > 1\\ 0, \left\|r_{BMU} - r_{SBMU}\right\| \le 1\end{cases}$$

де r_{BMU} – координати нейрона найкращого наближення, r_{SBMU} – координати *другого нейрона найкращого наближення* (*SBMU, second best matching unit*).

Зазначимо, що другим нейроном найкращого наближення для деякого вхідного вектора є нейрон, ваговий вектор якого є наступним за близькістю до нейрона-переможця. Іншими словами, нейрон SBMU стане BMU, якщо оригінальний BMU вилучити із ґратки. Значення параметра TE у досліджених нейромережах на тестовій множині прикладів подано на рис. 8.

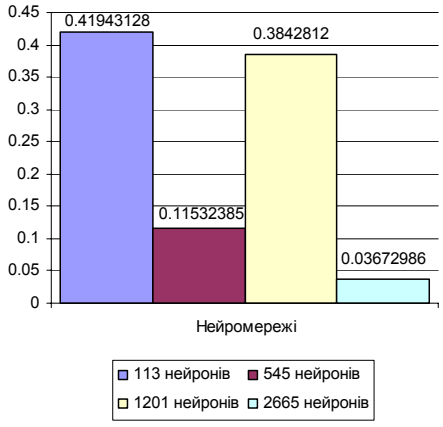

Рис. 8 – Значення параметру TE на тестових прикладах

Спостерігається покращення впорядкованості нейронів із збільшенням їхньої кількості у нейромережі. Найцікавішими можна вважати результати, які отримані аналізом структури нав-

чальних та тестових прикладів навченою нейромережею 2665. Значення топографічної помилки цієї нейромережі найнижче як для навчальної, так й тестової множин прикладів. Це свідчить про хорошу впорядкованість нейронів у ґратці, тобто, нейрони з близькими ваговими векторами розташовані близько між собою.

З огляду на високий рівень топологічного впорядкування нейромережі 2665 та беручи до уваги промарковані карти з рис. 5, 6 можна стверджувати, що значення атрибуту прийняття рішень, які встановлені суб'єктивно є хибними для певної кількості даних. Для виявлення таких хибно класифікованих даних можна скористатись „природним" розбиттям множини прикладів на кластери.

За значенням параметра PSR оцінено успішність розпізнавання хворого пацієнта як відсоток вхідних прикладів, правильно класифікований нейромережею. Обчислення цього параметру виконано порівнянням результатів класифікації, отриманих мережею із значеннями атрибуту прийняття рішення цих прикладів у досліджуваній системі прийняття рішень. Для обчислення значень параметра PSR використано класифікатор, запропонований авторами у праці [8].

Для побудови класифікатора зроблено припущення, що нейрони навченої мережі Кохонена здебільшого реагуватимуть на вхідні приклади з того кластера, на класифікації яких вони навчені. Отже, можна встановити вектори, на які повинен реагувати кожний нейрон. Якщо нейрон реагує на вектор зі свого кластера, то вважаємо реагування успішним, інакше – ні. Обчисленням кількості успішних та неуспішних реагувань кожного нейрона можна визначити його успішність. Також можна встановити успішність розпізнавання кожним нейроном як відсоток реагувань на певних пацієнтів за формулою

$$P_i = \frac{a_i}{a_i + b_i} \times 100\%,$$

де i – номер нейрона, a_i – кількість реагувань на пацієнтів, за розпізнавання яких відповідає нейрон i, b_i – кількість реагувань на пацієнтів, за розпізнавання який нейрон i не відповідає. Загальна успішність нейромережі Кохонена обчислюється як зважена сума успішностей усіх її нейронів.

На рис. 9 зображено залежність значень параметру PSR, отримане для досліджуваних нейромереж на тестовій множині прикладів, від розміру ґратки нейронів.

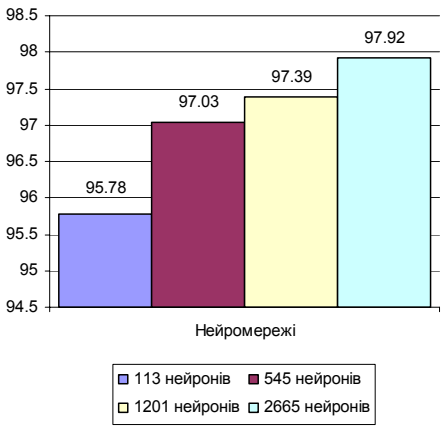

Рис. 9 – Успішність класифікації у відсотках на тестових прикладах

Свідченням кращої успішності є більше значення параметру PSR. Очевидно, що нейромережі із більшою кількістю нейронів краще класифікують дані. Водночас зазначимо, що нейромережу навчають на даних, які надано експертом-людиною. Ці дані містять помилки, викликані тим, що певна група пацієнтів класифіковані некоректно. Тому, попри високу успішність навчання, якість класифікації навченою нейромережею міститиме неявні похибки, що зумовлені природою навчальних даних. Це означає, що значення атрибуту прийняття рішень на прикладах, близьких у сенсі введеної відстані, може мати різне значення. Для виявлення і для їхнього виявлення таких неузгодженостей потрібно аналізувати на вхідні дані. Використання нейромережі SOM з метою кластеризації дає можливість краще вивчити структурні особливості аналізованих даних.

Аналіз атрибутів прийняття рішень для даних, які відповідають пацієнтам з одного кластеру, може виявити некоректні значення атрибуту прийняття рішень.

ВИСНОВКИ

Аналіз результатів експериментів за допомогою нейромереж, що самоорганізуються, дали змогу вдосконалити підходи до візуалізації результатів застосування цих нейромереж для класифікації даних. Отримані результати дозволили здійснити якісні висновки про результати застосування навченої нейромережі Кохонена до вирішення задач діагностування у формулюванні її як проблеми класифікації.

Головним результатом проведених досліджень вважаємо використання сукупності візуальних методів аналізу даних на предмет встановлення прихованих закономірностей у групах прикладів. Проведені дослідження показали можливість об'єднання методик,

основаних на різних підходах до аналізу таких закономірностей – класифікації та кластеризації. Таке поєднання дало змогу встановити певні прояви природної структуризації даних та можливості локалізації груп прикладів з метою виявлення в них суперечностей, викликаних суб'єктивною складовою у вирішенні задачі прийняття рішень.

Наступний етап досліджень вбачаємо у застосуванні кластерних технологій до вже навченої нейромережі Кохонена з метою об'єднання дрібних кластерів. Це дозволить полегшити аналіз даних з метою виявлення хибно класифікованих прикладів. Також важливим напрямком досліджень автори вважають побудову методів для автоматизації виділення точних границь кластерів на візуальних зображеннях карт висот.

ВИКОРИСТАНІ ДЖЕРЕЛА

[1] Czichosz P. Systemy uczace sie. – Wydawnictwa Naukowo-Techniczne, Warszawa, 2000.

[2] J. Komorowski, Z. Pawlak, L. Polkowski and A. Skowron (1999). Rough sets: A tutorial. In: S.K. Pal and A. Skowron (eds.), Rough fuzzy hybridization: A new trend in decision-making, Springer-Verlag, Singapore, pp. 3-98.

[3] Нікольський Ю.В., Пасічник В.В., Щербина Ю.М. Дискретна математика. – К.: Видавнича група BHV, 2007. – 368 с.

[4] Jiawei Han, Micheline Kamber. Data Mining: Concepts and Techniques. – Morgan Kaufmann Publishers, 2001.

[5] L.Kaufman, P.J.Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. – New York: John Wiley & Sons,1990.

[6] Нікольський Ю.В. Застосування методів кластерного аналізу при побудові класифікуючих правил в задачі прийняття рішень // Вісник Національного університету "Львівська політехніка", Інформаційні системи та мережі, 2003, № 489. – С.213-223.

[7] Dunhan M.H. Data Mining Introductory and Advanced Topics. – Prentice Hall, 2003,

[8] Годич О.В, Нікольський Ю.В., Пасічник В.В., Щербина Ю.М. Дослідження ефективності алгоритмів навчання мереж Кохонена. // Управляющие системы и машины, №2, 2006, с.63-80.

[9] Matti Pöllä, Timo Honkela, Henrik Bruun. Analysis of Interdisciplinary Text Corpora, Proceedings of the 12th Finnish Artificial Intelligence Conference STeP 2006, Helsinki University of Technology, Finland, October 26-27, 2006, – pp. 17-22

[10] Henrik Bruun, Sampsa Laine. Using the Self-Organizing Map for Measuring Interdisciplinary Research, Proceedings of the 12th Finnish Artificial Intelligence Conference STeP 2006, Helsinki University of Technology, Finland, October 26-27, 2006, – pp. 1-10.

[11] Jorma Laaksonen, Ville Viitaniemi. Emergence of ontological relations from visual data with Self-Organizing Maps, Proceedings of the 12th Finnish Artificial Intelligence Conference STeP 2006, Helsinki University of Technology, Finland, October 26-27, 2006, – pp. 31-38.

[12] M. Sirola, G. Lampi, J. Parviainen. SOM based decision support in failure manage-ment. International Journal of Computing, 4(3), 2005. – pp. 124-130.

[13] Joseph A. Cruz, David S. Wishart. Applications of Machine Learning in Cancer Prediction and Prognosis, Cancer Informatics 2, 2006. – pp. 59-78.

[14] Нікольський Ю.В., Щербина Ю.М., Якимечко Р.Я. Дерева прийняття рішень та їх застосування для прогнозування діагнозу у медицині // Вісн. Львів. ун-ту. Сер. прикл. мат. та інформ., 2003. – Вип. 4. – С. 191-211.

[15] Годич О.В., Нікольський Ю.В., Щербина Ю.М. Застосування штучної нейронної мережі типу SOM для розв'язування задачі діагностування // Вісник Національного університету "Львівська політехніка", 2002. – № 464. – С. 31-43.

[16] A. Ultsch. Self-Organizing Neural Networks for Knowledge Akquisition. In Proc. of the 10th ECAI, Vienna, Austria, 1992, – pp. 208-210.

[17] Si J., Lin S., Vuong M.-A. Dynamic Topology Representing Networks // Neural Networks, – 13 – 2000. – pp. 617-627.

Олесь Годич *закінчив факультет прикладної математики та інформатики Львівського національного університету ім. І. Франка у 2001 році. З 2005 року працює асистентом кафедри дискретного аналізу та інтелектуальних систем університету ім. І. Франка, на базі якої завершує дисертаційну роботу присвячену питанням кластеризації даних та аналізу часових рядів із використанням штучних нейронних мереж та суміжних технологій. У сферу його інтересів входять технології паралельних та розподілених обчислень, кластерний аналіз та*

розпізнавання образів, спеціалізовані мови програмування.

Юрій Нікольський *1975 року закінчив факультет автоматики Львівського політехнічного інституту. З 1975 по 1998 рік працював на факультеті прикладної математики та інформатики Львівського національного університету ім. І.Франка асистентом та доцентом, кандидат фізико-математичних наук. З 1998 року доцент кафедри „Інформаційні системи та мережі" Інституту комп'ютерних наук та інформаційних технологій Національного університету „Львівська політехніка". Викладає дисципліни, пов'язані із проблематикою дискретної математики та її методів у застосуванні до проблем та задач штучного інтелекту. Відмінник освіти України. Коло наукових інтересів охоплює проблематику штучного інтелекту, аналітичної обробки інформації, інженерії знань, прийняття рішень в умовах невизначеності, методи оптипізації*

Володимир Пасічник *– доктор технічних наук, професор. Народився 1956 року в родині сільських вчителів на Львівщині. Творчий шлях розпочав 1978 року після закінчення з відзнакою Львівського політехнічного інституту. У якому пройшов шлях від асистента до завідувача кафедри. Вихованець наукової школи Інститут кібернетики ім. В.М.Глушкова. Кандидат фізико-математичних наук (1982 рік, наукова спеціальність „Математична кібернетика") та доктор технічних наук (1994 рік, наукова спеціальність „Теоретичні основи інформатики"). Учасник та керівник багатьох міжнародних наукових проектів та перспективних науково-дослідних розробок. Відмінник освіти України, голова Львівської асоціації інформатиків. Наукові інтереси – інформаційне моделювання, системи баз даних та знань, розподілені інформаційні системи та технології.*

Юрій Щербина, *кандидат фізико-математичних наук, доцент кафедри дискретного аналізу та інтелектуальних систем Львівського національного університету ім. І.Франка., автор 64 наукових праць, 16 науково-методичних посібників та двох підручників з дискретної математики, написаних у співавторстві з Ю.Нікольським та В.Пасічником. Закінчив Львівський університет 1971 року. Читає лекції з курсів „Дискретна математика", „Методи оптимізації", „Нейронні мережі , генетичні алгоритми та розмиті системи". Коло наукових інтересів – дискретний аналіз, штучний інтелект, методи оптимізації.*

ANALYSIS OF THE MEDICAL DATA STRUCTURE USING SELF-ORGANISING MAPS

O.V. Hodych [2], Yu.V. Nikolsky [1], V.V. Pasichnyk [1], Yu.M. Scherbyna [2]

[1] National University „Lviv Polytechnics”
[2] Ivan Franko National Univerrsity

Abstract: *In this article the authors discuss several approaches to high dimensional data structure analysis using Self-Organising Maps. The describe approaches utilise graphical images for the purpose of data structure interpretation. The evaluation of the discussed techniques has been performed using the real medical data from cardiology. The research, results of which are outlined in this paper, is a continuation of the earlier work related to the analysis of the same medical data. It is envisaged that results obtained in this and earlier research work will form a foundation for creation of a robust technology to be used for automation of diagnostic tasks in medicine.*

Keywords: diagnostics, clustering, classification, artificial neural networks, data visualisation.

1. PAPER SIZE

In this paper authors discuss several approaches to retrieve some qualitative information about the structure of high dimensional data using its visual interpretation. The data used in this research is the medical data, which contains information about 3532 patients who may have some cardiological illness. This data has been used in the past research and the outlined in this paper work is based on previously obtained results.

The data forms a decision making table $B = (Z, A \cup \{d\})$ [2]. Symbol Z denotes a set of all objects (patients) being analysed; $A = (a_1, a_2, \ldots, a_m)$ is a row of attributes, which are the symptoms and results of analyses in case of cardiological data; each object from Z has a corresponding row of attributes, which can be represented as $a : Z \to V_a$, where V_a is a domain of attribute $a \in A$; d – is a decision making attribute. By removing a decision making attribute we obtain a table $B = (Z, A)$, which is called *information system*.

In our case the attribute domains are defined as follows: $V_{a_i} = \{0; 1\}$ $(i = \overline{1, m})$, $V_d = \{0; 1\}$. In other words, both patient attributes and a corresponding decision making attribute can be either 1 or 0. The information system is used for the purpose of analysis.

One of the widely used approaches to analyse data structures is clustering. Clustering methods provide an insight into the similarity between objects from the information system. The core idea behind clustering is grouping of objects based on their similarity. Many different definitions of a cluster have the following in common:

- Cluster – this is a set of similar objects; objects that exist in different clusters are considered to be not similar or different.
- Distance between objects inside one cluster is smaller that between objects from different clusters.

When applying clustering methods the following should be considered:

- In most cases there is no a priori knowledge about the number of clusters.
- There is no a priori knowledge as to how clusters can be effectively built.

The result of applying a clustering method to any data is a set of clusters $K = \{K_1, K_2, ..., K_l\}$, where $l \le n$ is a method parameter, which limits the number of possible clusters in the dataset of n samples.

There are many approaches to implement clustering. One of the most popular is hierarchical clustering, which consists of divisive and agglomerative methods [4]. The visualisation approach for representing results of hierarchical clustering utilises dendrogram – a tree diagram, which illustrates the arrangements of the clusters. An example of a simple dendrogram is depicted in Fig. 1.

Another effective clustering technology is Self-Organising Maps (SOM). Its core idea is in building a correspondence between high dimensional samples from an information system and neurons in one or two dimensional lattice. SOM neural network builds this relationship in an adaptive fashion by the means of learning. One of the key benefits of using SOM is

its ability to determine "natural" clusters. However, this ability is very dependent on many factors such as the size of a lattice and learning parameters. Earlier research related to the aforementioned medical data using SOM, which was conducted by the same authors, is described in [8] and [15]. The main purpose of utilising SOM in that research was building a more reliable and robust data model, which would have a low information/noise ratio. The produced model was successfully used for development of a classifier for providing a decision making support in diagnostics. Later is has become apparent that the original data contained erroneously assigned decision making attributes, and therefore the classifier was providing incorrect decisions despite its good performance on the original dataset. In order to resolve this problem the original data used for building a classifier needs to be corrected. Unfortunately, this is not a trivial task as it is not known what patients have incorrectly assigned decision making attribute. As the result, it has been decided to analyse the underlying structure of the original medical data in order to determine natural groupings between patients without the use of the decision making attribute.

The analysis of the information system B has been undertaken in the following ways:

- SOM neural network was trained on information system B.
- Trained SOM was used for building a map of heights utilising U-Matrix algorithm [16].
- Trained SOM was used for building a frequency map, where frequency is the number of times each individual neuron reacted to an input sample.
- Trained SOM was used to depict a distribution of neurons in the lattice based on the original decision making attribute assignments.

As mentioned earlier, one of the factors that affect SOM performance is the size of its lattice (i.e. number of neurons). In order to identify how does it affect SOM capabilities to visualise data structure several neural networks of different size have been used. The outlined in this paper results pertain to four neural networks with 113, 545, 1201 and 2665 neurons.

The first SOM had 113 neurons. The U-Matrix for this SOM did not provide a good visualisation and it is difficult to state how many and what is the location of clusters. The frequency map provided the information about the most active neurons, which were depicted with larger circles (about 10-15 of them), and the number of their reactions – labels above the circles. The maps, which correspond to a decision making attribute indicating presence of illness, were concentrated in the central part of the lattice with a little number of outliers.

The SOM with 545 neurons produced a slightly better result at depicting U-Matrix. The frequency map indicated a similar to SOM 113 result. Specifically, it contained approximately the same number of the most active neurons. It is important to note that the most active neurons are, in most cases, the centres of clusters. The decision making attribute based marking was similar to the maps produced for SOM 113, however the number of outliers is greater and, as it will be shown later, this tendency gets stronger with increase in number of neurons.

The last two SOM contained 1201 and 2665 neurons respectively. The results for SOM with 2665 neurons are depicted in Fig. 3-6. Both SOM provided a good visualisation of the data clusters with a slight advantage in case of SOM 2665. The number of the most active neurons in all four SOM is approximately the same. More interesting is the fact that the frequency of the most active neurons in all case is the same (e.g. neurons with 125, 92, 69, 88, 86, 49, 40, 39 reactions are present in all four maps). The decision making attribute based markings illustrate an even greater number of outliers and it becomes apparent that there is no concentration of neurons representing ill patients in the central part of a lattice. This leads to a conclusion that there is a discrepancy between natural grouping and decision making attribute grouping. In order to ensure the robustness of the trained SOM and therefore of the produced results, their evaluation has been undertaken utilising popular SOM performance indicators such as Mean Square Error (MSE) and Topological Error (TE). The results of this evaluation are depicted in Fig. 7 for MSE and Fig. 8 for TE. As can be observed, both MSE and TE provided a better result for SOM with larger lattices. In addition, the same classifier as discussed above was used to check classification capabilities of the trained SOM. The result of this evaluation is depicted in Fig. 9 showing consistently better results for SOM with larger lattices.

The obtained results strongly suggest that the structure of the medical data being analysed indeed contains discrepancy with the human assigned decision making attributes. The next step in this research is to identify rules, which would allow determining of the actual patients with incorrectly assigned attributes. The presented in this paper visualisation techniques, do not identify the exact borders of clusters. This would be one of the main tasks before identifying the correct decision making rules.

List of figures used in article: