МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ імені І.І.МЕЧНИКОВА

(повне найменування вищого навчального закладу)

Інститут математики, економіки та механіки

(повне найменування інституту, назва факультету (відділення))

Кафедра математичного забезпечення комп'ютерних систем

(повна назва кафедри (предметної, циклової комісії))

Дипломна робота

бакалавра

(освітньо-кваліфікаційний рівень)

на тему Метрична класифікація рівня розвитку інформаційних технологій на підставі статистичних даних

Metric classification of information technologies development level on the basis of statistical data

Метрическая классификация уровня развития информационных технологий на основе статистических данных

Виконав: студент 4 курсу, групи 1
напряму підготовки (спеціальності)
6.050102 Комп'ютерна інженерія
(шифр і назва напряму підготовки, спеціальності)
Кривонос В. О.
(прізвище та ініціали)
Керівник доц. Шпінарева І.М.
(прізвище та ініціали)
Рецензент доц. Мазурок І.Є.
(прізвище та ініціали)

Рекомендовано до захисту:

Протокол засідання кафедри

№_____ від «___»_____р.

Завідувач кафедри

_____ _Є.В. Малахов_

(підпис) (прізвище, ініціали)

Захищено на засіданні ЕК № ___

протокол №___ від «___» _____р.

Оцінка _____

(за 4-х бальною шкалою, за шкалою ECTS, бал.)

Голова ЕК

_____ _О.О. Арсірій_

(підпис) (прізвище, ініціали)

Одеса – 2017

АННОТАЦИЯ

Задача нахождения наиболее близкого объекта из некоторого класса к определенному заданному объекту другого класса является актуальной для многих прикладных областей. Сходство и различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними.

Актуальность дипломной работы заключается в необходимости анализа метрических методов классификации, и разработки программного обеспечения, которое осуществляет классификацию стран, разделяя их на существующие классы: высокий, средний, низкий.

Цель работы – создание инструментария автоматической классификации статистических данных.

Для достижения поставленной цели решена задача предварительного анализа выборки. Далее проведена экспертная классификация выборки и анализ степени важности параметров. Также проведена нормализация выборки. Следующим этапом выступает реализация модифицированного алгоритма KNN. Далее реализована система весовых коэффициентов для входных параметров и уже классифицированных объектов. Для удобства работы с системой создан графический интерфейс взаимодействия с пользователем. Для оценки качества работы алгоритма создана системы оценки классификации на основе результатов классификации контрольной выборки и обучающей выборки.

В результате проведеной работы было получено увеличение качества классификации контрольной выборки с 89% до 96%.

АНОТАЦІЯ

Тема дипломної роботи «Метрична класифікація рівня розвитку інформаційних технологій на підставі статистичних даних».

Актуальність дипломної роботи полягає в необхідності аналізу метричних методів класифікації, та розробки програмного забезпечення, яке здійснює класифікацію країн, поділяючи їх на існуючі класи: високий, середній, низький.

Об'єкт дослідження – метричні алгоритми класифікації та їх застосування для класифікації рівня розвитку інформаційних технологій країн світу.

Мета роботи – створення інструментарію автоматичної класифікації статистичних даних.

Для досягнення поставленої мети вирішена задача попереднього аналізу вибірки. Далі проведена експертна класифікація вибірки і аналіз ступеня важливості параметрів. Також проведена нормалізація вибірки. Наступним етапом виступає реалізація модифікованого алгоритму KNN. Далі реалізована система вагових коефіцієнтів для вхідних параметрів і вже класифікованих об'єктів. Для зручності роботи з системою створений графічний інтерфейс взаємодії з користувачем. Для оцінки якості роботи алгоритму створена система оцінки класифікації на основі результатів класифікації контрольної вибірки і навчальної вибірки.

В результаті проведеної роботи було отримано збільшення якості класифікації контрольної вибірки з 89% до 96%.

ABSTRACT

Theme of the diploma work is "Metric classification of the level of development of information technology on the basis of statistical data."

The relevance of diploma work is the need to analyze metric classification methods, and develop a software, which classifies countries dividing them by existing classes: high, medium, low.

The object of research is the metric classification algorithms and their application for the classification of the information technologies development level of the countries of the world.

The purpose of the work is to create a toolkit for automatic classification of statistical data.

To achieve this goal, the task of preliminary analysis of the sample is solved. Next, an expert classification of the sample and an analysis of the importance of the parameters were carried out. Also, the sample was normalized. The next step is the implementation of the modified KNN algorithm. Next, a weighting system is implemented for input parameters and already classified objects. For the convenience of the system, a graphical user interface is created. To assess the quality of the algorithm, classification evaluation systems are created based on the results of the classification of the control sample and the training sample.

As a result of the work, an increase in the quality of the control sample classification was obtained from 89% to 96%.

ЗМІСТ

ВСТУП

На даний момент, обсяги інформації накопичені людством досягають досить великих розмірів, і існує необхідність різноманітної обробки цих даних. Необхідність класифікації даних, що надходять, досить висока в зв'язку з необхідністю пошуку інформації. Якщо інформації, що надходить, не класифікувати, то процес пошуку стає вкрай важкою задачею. Одним з ефективних практичних підходів до класифікації об'єктів є зіставлення аналізованого опису з кінцевим набором еталонних описів. Такі підходи називають метричними [1]. Основу метричного підходу становить побудова описів і синтез міри подібності для оцінки їх близькості.

Методи метричної класифікації вирішують проблему розподілу даних на певні класи, і необхідні для класифікації даних, представлених в еквівалентному цифровому вигляді. Кожен об'єкт, класифікується, в такому випадку представляється у вигляді набору значень. В основі методів метричної класифікації лежить припущення про схожість об'єктів за описом, іншими словами, якщо два об'єкти, представлені у вигляді набору значень, мають схожі параметри, система відносить такі об'єкти до одного класу.

Розглянемо задачу класифікації. Нехай в n-вимірному просторі ознак розташоване k-об'єктів з класу А. Об'єкт класу В також характеризується n-ознаками, але в загальному випадку значення деяких ознак може бути не визначени (не задани) або для однієї ознаки може бути задано кілька значень. Крім цього, слід враховувати актуальність (необхідність) кожної ознаки. Наприклад, ознака (характеристика) об'єкта В приймає значення X_1 і X_2, а актуальність (важливість) даної ознаки може бути визначена значенням шкали «важлива характеристика; скоріше важлива характеристика, ніж ні; характеристику можна не враховувати». Для об'єкта класу В необхідно визначити найбільш близький, з точки зору збігу значень по n-ознакам, об'єкт класу А. Шукана функція близькості між об'єктами класів В і А повинна

враховувати багатозначність у визначенні деяких ознак, а також актуальність кожної ознаки. Для визначення ступеня схожості використовуються різноманітні метрики. Метрика – функція визначення відстані між об'єктами.

Розглянута задача знаходження найбільш близького об'єкта з деякого класу до певного заданого об'єкта іншого класу є затребуваною для багатьох прикладних областей.

У дипломній роботі розглянемо методи метричної класифікації та обраним алгоритмом вирішимо задачі визначення рівня розвитку інформаційних технологій в країнах світу.

Для цього необхідно вирішити такі задачі:

– огляд метричних алгоритмів;
– реалізувати алгоритм класифікації k найближчих сусідів;
– реалізувати різні метрики;
– реалізувати різні види нормалізації даних;
– реалізувати систему вагових коефіцієнтів важливості параметрів;
– створити інтерфейс взаємодії з користувачем.

ВИСНОВКИ

У дипломній роботі створено інструментарій автоматичної класифікації даних. В якості предметної області виступає вибірка про рівень розвитку інформаційних технологій в країнах світу. В якості класів виступають такі рівні розвитку: високий, середній, низький.

Класифікації даних виконується за допомогою модифікованого алгоритму KNN. До нього додані наступні параметри:

– нормалізація. У реалізованій програмі присутня можливість вибору одного з двох варіантів нормалізації – стандартної і мінмаксної;

– коефіцієнти важливості параметрів. Реалізована можливість вибору коефіцієнтів важливості для кожного з параметрів предметної області, доступні значення від одиниці до ста. Коефіцієнти підібрані методом відпалу і, при запуску програми, налаштовані на отримання масимального результату класифікації;

– урахуванням відстані до сусідів, що беруть участь у виборі класу. Реалізовано вибір враховувати відстань чи ні;

– кількість сусідів, що беруть участь у виборі класу;

– метрики. Реалізовано три метрики доступних для вибору: Евклідова, Манхеттенська, Мінковського.

З отриманих даних випливає, що на контрольній вибірці введення модифікацій приводить до поліпшення якості та стабільності роботи алгоритму при різних k. Максимальна якість класифікації підвищилося з 89% до 96%. Мінімальна якість класифікації підвищилося з 65% до 76%. Середнє значення якості класифікації також підвищився з 74.6833% до 86.8%.

На згенерованій вибірці в 100 000 елементів алгоритм показує себе краще ніж на контрольній вибірці. Максимальна якість класифікації підвищилася з 99.301% до 99.661%. Мінімальна якість класифікації знизилася з 87% до 82%. Середнє значення якості класифікації підвищилося з

95.645% до 97.217%. На згенерованої вибірці, введення нормалізації знижує якість класифікації при $k < 3$. Однак в іншому модифікації є вкрай вдалими. На контрольній вибірці введення будь-якої модифікації показує себе хорошим чином підвищуючи якість класифікації, проте збільшуючи необхідне для максимальної якості класифікації значення k.

Таким чином, модифікований в даній роботі алгоритм KNN, в середньому показує результати на 10-12% вище ніж стандартний KNN.

ПЕРЕЛІК ПОСИЛАНЬ

1. Метод k ближайших соседей [Електронний ресурс] – Режим доступа: https://basegroup.ru/community/glossary/nearest-neighbor – 25.03.2017
2. Задачи классификации [Електронний ресурс] – Режим доступа: http://www.machinelearning.ru/wiki/index.php? Классификация – 24.03.2017
3. Воронцов К.В. Лекции по метрическим алгоритмам классификации [Електронний ресурс] – Режим доступа: http://www.ccas.ru/voron/download/MetricAlgs.pdf – 27.03.2017
4. Скользящий контроль [Електронний ресурс] – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Кросс-валидация – 23.03.2017
5. Алгоритм имитации отжига [Електронний ресурс] – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Алгоритм_имитации_отжига – 25.04.2017
6. Методы многомерных классификаций [Електронний ресурс] – Режим доступа: http://ecocyb.narod.ru/513/MSM/msm2_2.htm – 24.03.2017
7. Метрика Минковского [Електронний ресурс] – Режим доступа: http://hypercomplex.xpsweb.com/articles/510/ru/pdf/11-14.pdf – 24.04.2017
8. Алгоритм ближайшего соседа в решении задач классификации [Електронний ресурс] – Режим доступа: https://basegroup.ru/community/articles/knn – 30.04.2017

ДОДАТОК

Статистична вибірка рівня розвитку інформаційних технологій країн світу

Назва країни	Код країни	Кількість ПК на 1000 жителів	Кількість користувачів Інтернет на 1000 жителів	Кількість хостов Интернет на 100000 жителів	Річний дохід від сфери телекоммунікацій (в млн. дол. США)	Річні інвестиції в сферу телекоммунікацій (в млн. дол. США)
1	**2**	**3**	**4**	**5**	**6**	**7**
Буркина-Фасо	BFA	2	4	4	63	24
Камбоджа	KHM	2	2	6	19	22
Эритрея	ERI	2	3	25	19	28
Мали	MLI	2	3	2	92	18
Бенин	BEN	4	10	12	84	26
Нидерланды	NLD	4	3	4	84	21
Судан	SDN	6	9	0	182	128
Гвинея	GIN	6	5	5	29	1
Коморские о-ва	COM	6	6	1	10	4
Кения	KEN	6	13	26	606	45
Йемен	YEM	7	5	1	144	74
Замбия	ZMB	8	6	17	69	5
Алжир	DZA	8	16	3	362	96
Бангладеш	BGD	8	2	2	497	80
Кот-д'Ивуар	CIV	9	14	23	540	77
Албания	ALB	12	10	8	251	32
Бутан	BTN	14	20	134	9	3
Гамбия	GMB	14	19	42	30	4
Кыргызстан	KGZ	14	38	110	60	1
Азербайджан	AZE	15	43	7	102	29
Гондурас	HND	15	40	28	372	104
Армения	ARM	16	37	55	80	30
Шри-Ланка	LKA	17	13	10	335	82
Монголия	MNG	21	80	332	124	44
Сенегал	SEN	21	22	6	254	109
Габон	GAB	22	26	21	143	44
Джибути	DJI	22	10	100	33	14
Боливия	BOL	23	37	84	473	162
Куба	CUB	24	9	14	876	111
Нигер	NER	27	17	129	153	37
Сербия и Черногория	SCG	27	79	184	358	212
Того	TGO	32	42	2	42	30
Сальвадор	SLV	33	83	62	2	3
Парагвай	PRY	34	20	156	309	82
Оман	OMN	37	71	28	530	127
Тунис	TUN	40	64	3	876	306
Суринам	SUR	42	44	3	68	12
Турция	TUR	42	85	508	8	230
Иордания	JOR	45	81	57	828	149
Фиджи	FJI	51	67	60	71	38
Ямайка	JAM	53	228	56	548	136

Продовження таблиці

1	2	3	4	5	6	7
Мальдивские о-ва	MDV	70	53	185	65	8
Кабо-Верде	CPV	76	44	14	48	8
Тринидад и Тобаго	TTO	76	106	614	299	110
Барбадос	BRB	104	371	76	181	44
Белиз	BLZ	119	109	885	62	12
Бахрейн	BHR	154	216	192	505	66
Катар	QAT	156	199	31	557	103
Мальта	MLT	253	475	1779	257	57
Кипр	CYP	260	337	779	429	110
Пакистан	PAK	4	8	10	1678	210
Хорватия	HRV	174	232	678	1240	182
Латвия	LVA	188	404	1779	249	40
Перу	PER	42	104	240	1395	175
Северные Марианские о-ва	NMP	7	6	1	1217	3869
Мозамбик	MOZ	20	33	12	1966	278
Таиланд	THA	39	96	164	4141	401
Венесуэла	VEN	60	60	137	2456	237
Филиппины	PHL	27	44	35	2952	697
Египет	EGY	29	44	5	2417	513
Украина	UKR	24	53	192	2340	773
Колумбия	COL	43	59	263	3684	1530
Южная Африка	ZAF	71	68	623	5339	712
Греция	GRC	79	150	1705	6820	1258
Аргентина	ARG	81	112	2007	3764	869
Российская Федерация	RUS	89	68	422	6956	1015
Иран	IRN	90	72	8	1715	1264
Румыния	ROM	97	184	218	1727	302
Венгрия	HUN	106	232	3578	4661	486
Литва	LTU	110	202	1926	663	101
ОАЭ	ARE	111	340	1390	2534	312
Саудовская Аравия	SAU	133	67	70	7278	1541
Португалия	PRT	134	257	2181	7880	1975
Чили	CHL	136	272	1376	2421	589
Польша	POL	142	232	2038	9670	2450
Эстония	EST	440	444	4741	551	41
Малайзия	MYS	167	344	429	4792	1009
Чешская респ.	CZE	179	308	2744	3999	1504
Словакия	SVK	236	256	2122	1381	345
Словения	SVN	325	401	2148	844	170
Индонезия	IDN	12	38	29	2167	1703
Индия	IND	7	17	8	7959	3512
Израиль	ISR	238	301	6439	4342	1441
Бельгия	BEL	318	386	2026	8196	1139
Ирландия	IRL	416	317	3992	4067	289
Люксембург	LUX	620	337	6249	413	145
Китай	CHN	28	63	13	55527	26782
Бразилия	BRA	74	102	1798	20428	5205

Продовження таблиці

1	2	3	4	5	6	7
Испания	ESP	195	239	2224	38610	5103
Италия	ITA	237	416	1149	35241	7289
Япония	JPN	382	483	10157	168914	19997
Австрия	AUT	396	462	7134	5663	905
Нигерия	NGA	407	526	11833	3159	301
Великобритания	GBR	412	592	5453	72836	13433
Франция	FRA	417	366	4012	39708	5472
Финляндия	FIN	441	534	24365	4992	730
Исландия	ISL	450	675	37897	174	37
Нидерландские Антильские о-ва	ANT	464	522	21627	13138	2633
Тайвань, китайская провинция	TWN	471	519	12286	10300	3552
Канада	CAN	482	555	10120	22727	3771
Германия	GER	485	400	3154	71011	5618
Бермудские о-ва	BMU	520	464	13468	88	25
Норвегия	NOR	525	346	12459	7253	2589
Гонконг	HKG	555	472	8693	6102	1034
КНДР	KOR	558	610	7976	24433	8033
Дания	DNK	574	541	23127	5494	850
Австралия	AUS	602	567	14281	14894	4663
Сингапур	SGP	617	509	11553	3349	433
Швеция	SWE	619	630	10507	7824	1482
США	USA	688	556	55778	295720	17633
Швейцария	CHE	742	398	7489	11043	1633