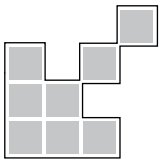

ТЕХНОЛОГИЯ ПРОИЗВОДСТВА. ПРИБОРЫ И УСТРОЙСТВА

УДК 004.031.2;025.4

Ю.А. Радионова

ИНСТРУМЕНТАРИЙ ОЦЕНКИ ЭФФЕКТИВНОСТИ МЕТОДОВ АВТОМАТИЧЕСКИХ КЛАСТЕРИЗАЦИЙ

***Радионова Юлия Александровна**, аспирант кафедры «Информационные системы» Ульяновского государственного технического университета. Ведущий инженер-программист ФНПЦ ОАО «НПО «Марс». Сфера научных интересов — автоматизированные системы документооборота, организация хранилищ технической документации. [e-mail: julia-owl@mail.ru].*

Аннотация

В статье рассматривается инструментарий для автоматизации процессов экспертной классификации электронной документации, сравнения результатов экспертной классификации и автоматических кластеризаций методами Кохонена и FCM, а также построения оценочных функций по полученным результатам сравнений.

Ключевые слова: программное приложение, автоматизация классификации, оценка автоматической кластеризации, экспертная классификация, параметры кластеризации.

Abstract

The article deals with tools to ensure automation of expert classification processes for e-documents, comparison of results of expert classification and automatic clustering using Kohonen and FCM methods as well as creation of evaluation functions as a result of comparisons.

Key words: application, computer-aided classification, automatic clustering evaluation, expert classification, clustering parameters.

Введение

Для оценки эффективности методов автоматических кластеризаций разработана методика сравнения данных автоматических кластеризаций с данными классификации, проведенной архивариусом-экспертом, методика построения оценочной функции, а также план экспериментов [1].

В связи с достаточно большими объемами массива классифицируемой документации и большой размерностью получаемых матриц соответствия необходимо разработать инструментарий, позволяющий автоматизировать процессы экспертной классификации и алгоритмы сравнения данных и построения оценочных функций.

Для автоматизации процессов экспертной классификации, а также построения матриц соответствия и оценочных функций было разработано приложение «Классификация документов электронного архива».

Приложение разработано в среде Borland Delphi 7.0 и использует базу данных в формате MS SQL Server 2000.

Описание программного приложения

Главное окно программы, представленное на рисунке 1, содержит строку меню и панель с вкладками, на каждой из которых реализована та или иная функция приложения.

Для подключения к базе данных необходимо ввести наименование SQL-сервера, установленного на данном компьютере, справа сверху на вкладке «Список документов» и нажать кнопку «Подключить».

В левой части вкладки «Список документов» представлен список документов, которые будут участвовать в экспертной и автоматических классификациях. Список становится видимым при подключении к базе данных.

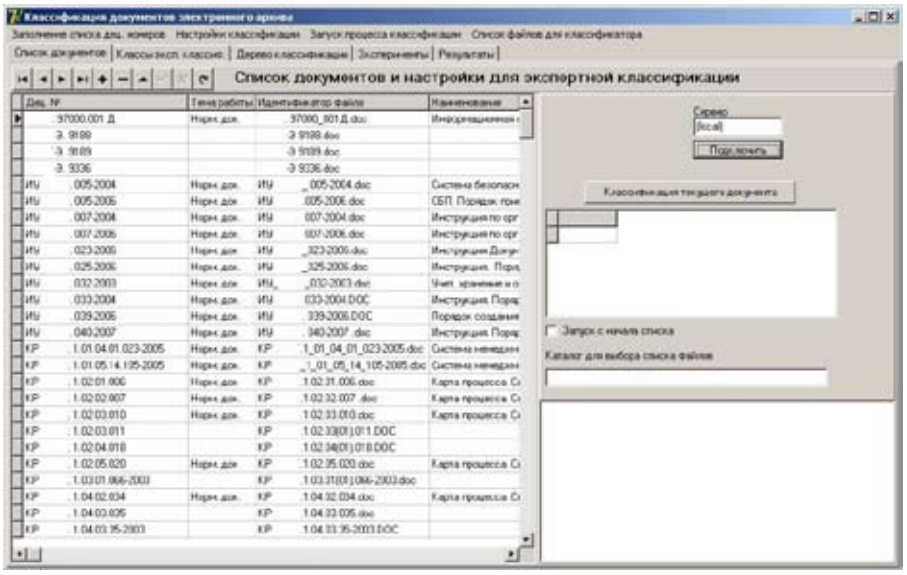

Рис. 1. Главное окно приложения

Рис. 2. Пункт главного меню «Заполнение списка дец. номеров»

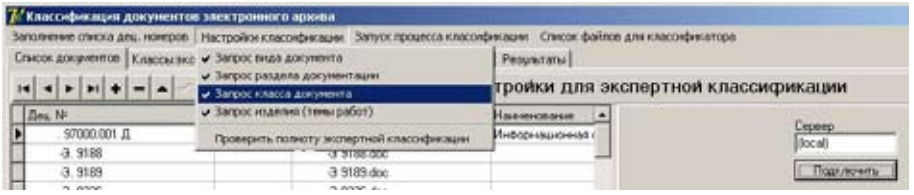

Рис. 3. Пункт главного меню «Настройки классификации»

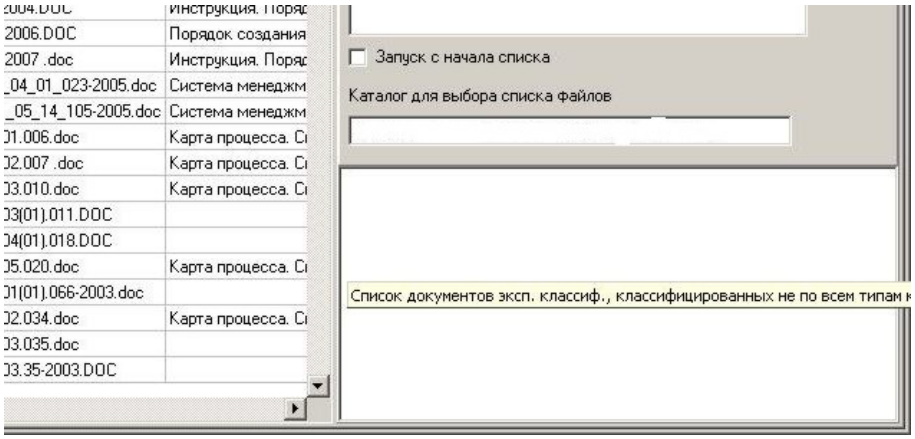

Рис. 4. Пункт «Список документов с неполной классификацией»

Функции экспертного классификатора

Выборка списка документов для экспертной классификации возможна как из произвольного каталога файлов, так и из электронной картотеки архива машинных носителей ФНПЦ ОАО «НПО «Марс». Для этой функции служит пункт главного меню «Заполнение списка дец. номеров» (рис. 2).

Выбор подпункта «Из картотеки» позволяет заполнить список документов из электронной картотеки архива. При выборе подпункта «Из каталога файлов» список документов заполняется исходя из набора файлов каталога, который указан в поле «Каталог для выбора списка файлов» в нижней правой части вкладки «Список документов». При этом производится выделение децимального номера документа из идентификатора файла и приведение его к стандартному виду [2]. Для упрощения процесса дальнейшей классификации документов существует возможность внести в список темы, в рамках которых разработаны занесенные в список документы, подпункт меню «Заполнить темы». Темы работ выбираются из картотеки архива.

Пункт «Настройки классификации» (рис. 3) служит для уточнения параметров экспертной классификации, а именно, по каким из четырех типов будет проходить процесс классификации.

Также в данном пункте можно проверить полноту проведенной экспертной классификации — все

ли документы, указанные в списке, классифицированы по выбранным типам классификации. Список документов с неполной классификацией приводится в правой нижней части вкладки «Список документов» (рис. 4).

В пункте «Запуск процесса классификации» можно запустить классификацию как всего списка документов, так и отдельного текущего документа (рис. 5).

Если в поле «Запуск с начала списка» (правая средняя часть вкладки) стоит отметка, то при запуске классификации всего списка процесс начинается с начала списка, иначе процесс будет начат с той строки списка, на которой стоит курсор.

При выборе пункта «Список файлов для классификатора» можно сформировать каталог с файлами для процесса классификации, запрос на наименование каталога будет выдан оператору в процессе работы. Файлы можно выбирать из электронной картотеки архива — подпункт «Из картотеки», при этом выбираются файлы только действующих документов и формата MS Word. Также файлы можно выбирать из произвольного каталога, при этом каталог-источник может иметь иерархическую структуру с неограниченным числом уровней. Запрос на наименование каталога-источника будет также сформирован в процессе работы приложения.

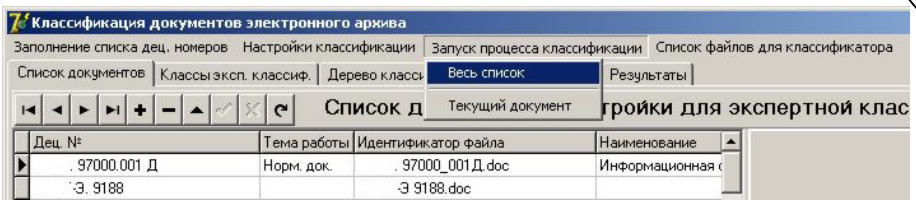

Рис. 5. Пункт главного меню «Запуск процесса классификации»

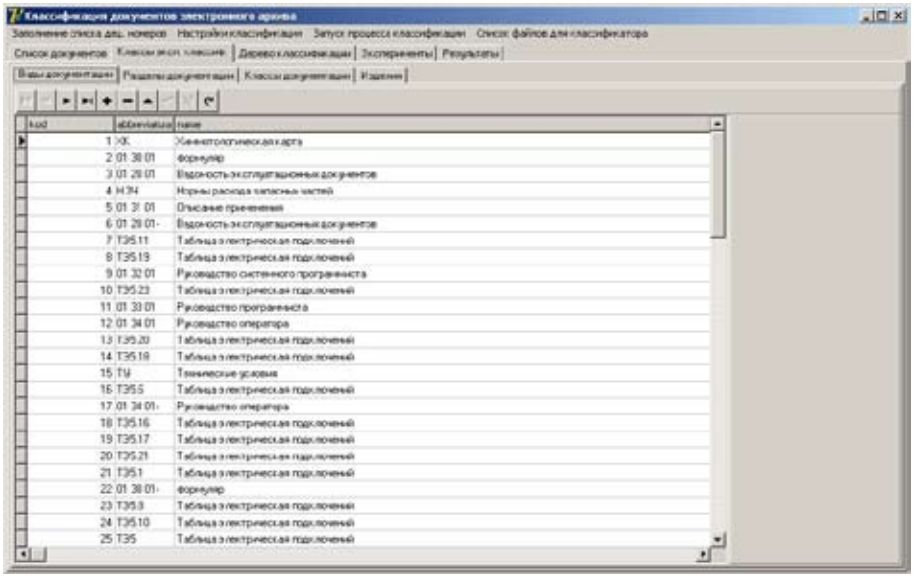

Рис. 6. Вкладка «Классы экспертного классификатора»

Рис. 7. Классификация текущего документа

Экспертная классификация документов

Процесс экспертной классификации проходит интерактивно с участием оператора. Документы классифицируются по четырем признакам: виду документа, разделу документации, классу документации и тематике работ. Для каждого типа классификации в базе данных содержится справочник классов, используемых в процессе классификации. Справочник каждого типа в приложении можно просматривать и редактировать на вкладке «Классы эксп. классиф.» (рис. 6).

Алгоритм классификации основан на анализе децимального номера документа и подробно описан в работе [2]. В процессе классификации программа проверяет наличие соответствующих классов в справочниках классификатора. Если класс найден, то он присваивается документу по данному типу классификации. Иначе программа формирует запрос к оператору на ввод нового класса данного типа.

Работу алгоритма классификации можно остановить в любое время, а также продолжить с любого документа списка.

Результат классификации можно просмотреть как для каждого отдельного документа — по двойному клику на списке в таблице в правой средней части вкладки «Список документов» (рис. 7), так и для всей совокупности документов — на вкладке «Дерево классификации» (рис. 8). Для первоначальной инициализации древовидной структуры необходимо нажать кнопку «Заполнить дерево».

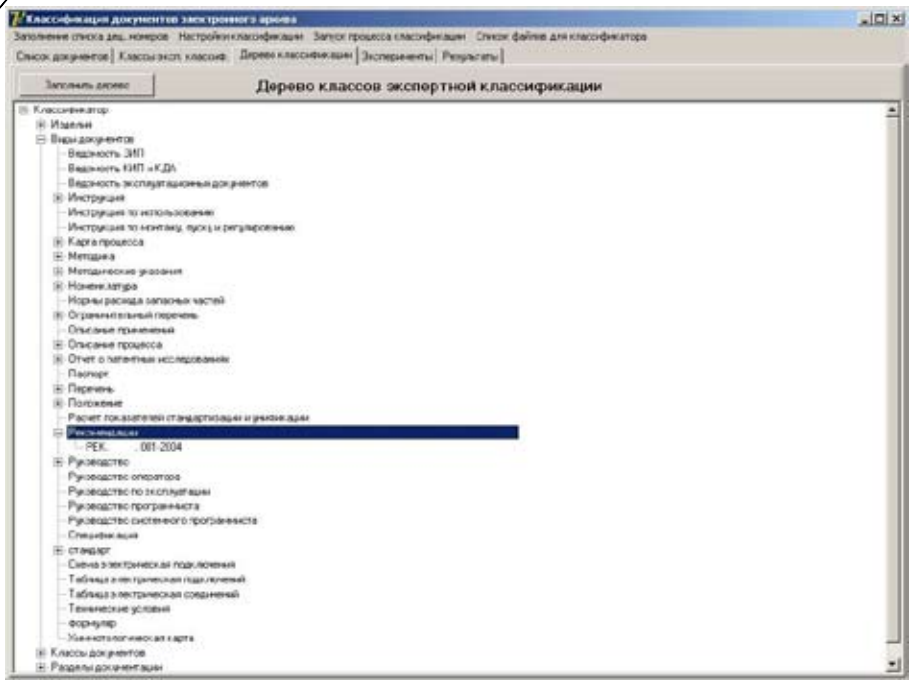

Рис. 8. Вкладка «Древовидная структура классификатора»

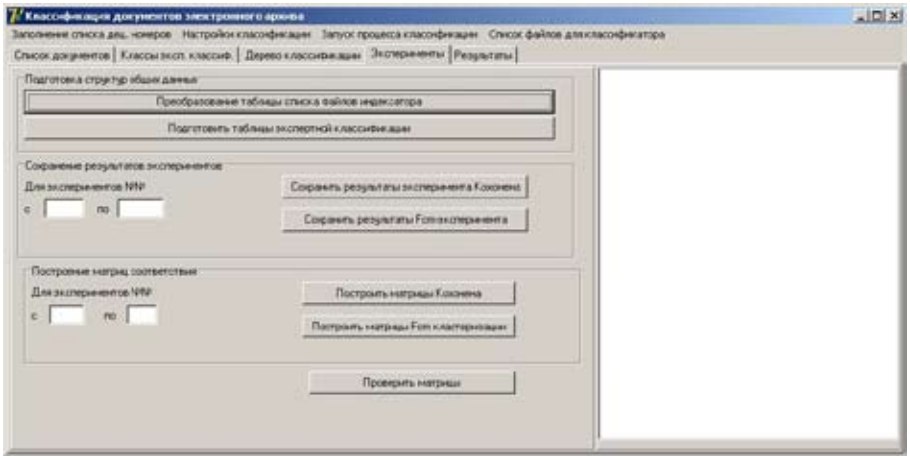

Рис. 9. Вкладка «Эксперименты»

- подготовка таблиц экспертной классификации, заключающуюся в объединении подобных классов в экспертном классификаторе, а также формировании общей таблицы кластеров для всех типов экспертных классификаций со сквозной нумерацией кластеров и таблицы соответствия кодов документов кодам кластеров экспертной классификации.

Второй этап «Сохранение результатов эксперимента» заключается в преобразовании структур сохраненных таблиц кластеризации Кохонена и FCM-кластеризации. Таким образом, получаем таблицы кластеров и распределения документов по кластерам для каждого эксперимента алгоритмов Кохонена и FCM со сходной структурой.

Третий этап «Построение матриц соответствия» заключается в построении для каждого типа экспертной классификации и каждого эксперимента автоматической кластеризации матрицы соответствия, строки которой содержат код кластера автоматической кластеризации, код кластера экспертной классификации, количество совпадающих документов в соответствующих кластерах и общее количество документов в каждом кластере экспертной и автоматической кластеризаций. Результатом выполнения третьего этапа являются таблицы:

resMatrixClass_documN
resMatrixDocumentsN
resMatrixIzdeliaN
resMatrixRazdelN
resFcmMatrixClass_documM
resFcmMatrixDocumentsM
resFcmMatrixIzdeliaM
resFcmMatrixRazdelM,

где N — эксперименты кластеризации Кохонена;

M — эксперименты FCM-кластеризации.

Процесс формирования матриц соответствия отражается в поле вывода информации в правой части вкладки.

Результаты экспериментов автоматических кластеризаций

На вкладке «Эксперименты» (рис. 9) расположен инструментарий для «подгонки» различных структур данных кластеризации Кохонена и FCM-кластеризации к единому виду для подготовки данных автоматических кластеризаций и экспертной классификации к сравнению.

Инструментарий для подготовки данных можно условно разбить на три части — этапа подготовки.

Первый этап «Подготовка структур общих данных» подразделяется на следующие операции:

- преобразование таблицы списка файлов индексатора, заключающееся в формировании таблиц соответствия кодов документов, используемых в экспертной классификации и автоматических кластеризациях;

Для контроля правильности работы алгоритма экспертной классификации, преобразования структур таблиц и построения матриц соответствия в приложении предусмотрены несколько точек контроля. Действие первой точки контроля реализовано на вкладке «Список документов» — проверка полноты проведения экспертной классификации. Вторая точка контроля — проверка матриц соответствия — заключается в подсчете количества документов каждого класса экспертной классификации, совпавших с документами различных кластеров автоматической кластеризации, и сравнении этого количества с общим количеством документов в классе (то же самое для каждого кластера автоматической кластеризации). Если обнаружено несовпадение, то это значит, что допущена ошибка в работе алгоритма, либо неверно (неполно) проведен процесс экспертной классификации, либо допущено нарушение структуры данных каких-либо таблиц базы данных.

На вкладке «Результаты» можно просмотреть значения оценочных функций для всех типов экспертной классификации и экспериментов автоматических кластеризаций, а также матрицы соответствия и списки документов, содержащихся в различных кластерах и совпавших для конкретных классов экспертной классификации и кластеров автоматической кластеризации (рис. 10).

В верхней левой части вкладки указываются номера экспериментов, для которых необходимо подсчитать оценочные функции, и соответствующей кнопкой выбирается вид автоматической кластеризации — Кохонена или FCM. Для каждого типа экспертной классификации формируется отдельный список значений оценочной функции. В заголовке списка указывается тип экспертной классификации и в скобках — количество классов данного типа. Каждая строка списка имеет вид:

$$N\,XXX(YY) - F,$$

где XXX — номер эксперимента автоматической кластеризации;

YY — количество кластеров в данном эксперименте;

F — значение оценочной функции.

Двойным кликом на строке списка в левой нижней части вкладки открывается матрица соответствия для данного эксперимента и типа экспертной классификации. В заголовке матрицы указываются номер эксперимента и тип классификации (на рисунке — это 007, Разделы документации). Каждая строка матрицы содержит информацию:

kod_exp — код класса экспертной классификации;

kod_avt — код кластера автоматической кластеризации;

count_eq — количество документов, совпавших для данных класса и кластера;

exp_all — общее количество документов экспертного класса;

avt_all — общее количество документов автоматического кластера.

Двойной клик на строке матрицы соответствий открывает в правой части вкладки списки документов, содержащихся в соответствующем кластере автоматической кластеризации; документов, содержащихся в классе экспертной классификации; а также для кластеризаций Кохонена список термов для текущего кластера. В правой нижней части вкладки открывается список документов, общих для данных кластера и класса. В заголовке списков указываются тип экспертной классификации и название текущего класса.

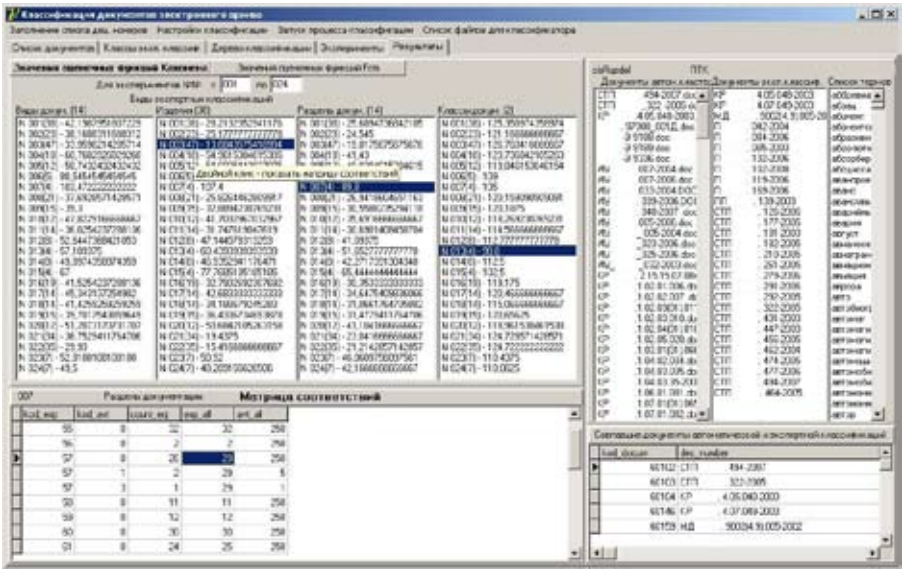

Рис. 10. Вкладка «Результаты экспериментов»

Оценка результатов экспериментов

Результаты экспертной классификации

Проведены два этапа экспериментов.

На первом этапе из архива электронной документации выбраны 65 документов преимущественно организационно-нормативного содержания. Проведена экспертная классификация по четырем признакам, и получено следующее количество классов:

- по виду документов — 16 классов;

- по тематике работ — 22 класса;
- по классу документации — 3 класса;
- по разделу документации — 22 класса.

На втором этапе из архива электронной документации выбраны 265 документов исключительно организационно-нормативного содержания. Также проведена экспертная классификация по четырем признакам, и получено следующее количество классов:

по виду документов — 14 классов;
по тематике работ — 38 классов;
по классу документации — 2 класса;
по разделу документации — 14 классов.

Нормировочные коэффициенты функции соответствия

В процессе проведения экспериментов была выявлена существенная зависимость значения целевой функции от количества классов экспертной классификации и кластеров автоматической кластеризации — чем больше количество кластеров, тем больше строк в матрице соответствия, следовательно, больше слагаемых при вычислении оценочной функции и больше значение самой функции. Например, динамика значений целевой функции для экспериментов Кохонена и видов документации в экспертной классификации по второй части плана экспериментов представлена на рисунке 11 (по горизонтальной оси — количество кластеров, по вертикальной — значения оценочной функции).

Таким образом, получаем, что значение целевой функции зависит больше от количества кластеров, чем от эффективности кластеризации. Для того чтобы убрать данную зависимость, эксперименты проводились с разными нормирующими коэффициентами целевой функции:

- $Nэ$ — количество классов экспертной классификации;
- Na — количество кластеров автоматической кластеризации;
- C — количество строк в текущей матрице соответствия.

Для проведения экспериментов использовались следующие нормирующие коэффициенты:

- $Nэ+Na$;
- $Nэ*Na$;
- C.

Примеры применения коэффициентов приведены на рисунках 12, 13, 14 соответственно.

Как видно из диаграмм, первые два значения коэффициента дают обратную зависимость — чем меньше количество классов и кластеров, тем хуже значение целевой функции — ярко выраженные пики на малых значе-

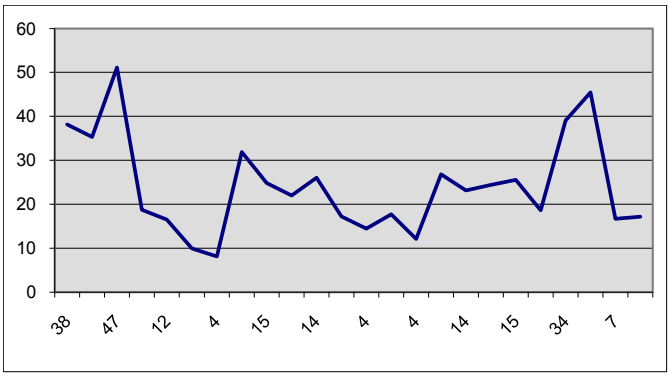

Рис. 11. Значения целевой функции для экспериментов Кохонена второй части

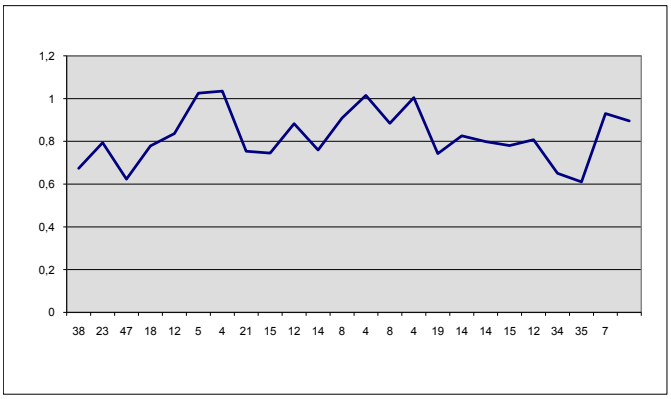

Рис. 12. Значения целевой функции с первым коэффициентом «$Nэ+Na$»

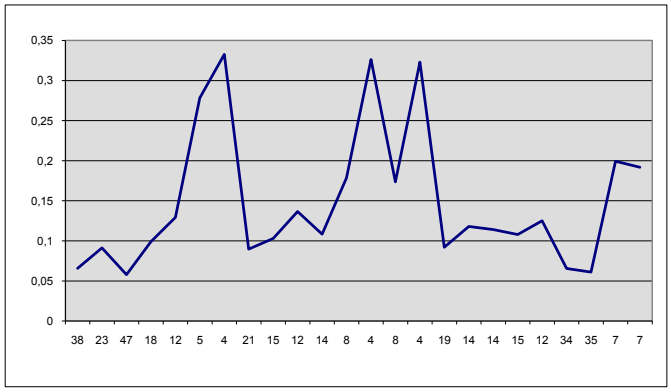

Рис. 13. Значения целевой функции со вторым коэффициентом «$Nэ*Na$»

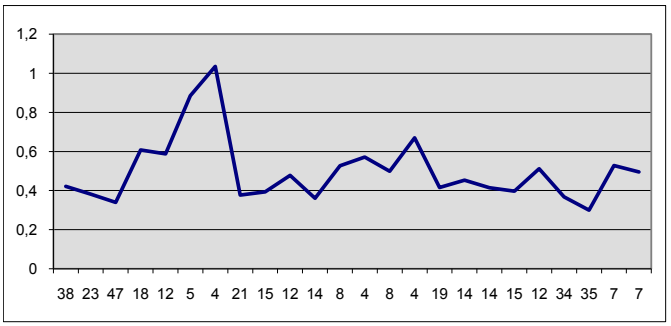

Рис. 14. Значения целевой функции с третьим коэффициентом «C»

Таблица 1

Параметры и значения целевой функции для алгоритма Кохонена второй части плана экспериментов

№ эксп.	Число нейронов	Радиус активации	Норма обучения	Множитель нормы обучения	Полученное количество кластеров	Значения целевой функции для типов экспертной классификации			
						виды документации (14)[2]	изделия (38)	разделы (14)	классы документации (2)
1	12	2	0,9	0,9	39	0,422	0,2821	0,2569	1,2536
2	5	2	0,5	0,5	23	0,3817	0,2518	0,2454	1,2117
3	8	2	0,7	0,7	47	0,3396	0,1368	0,1982	1,2676
4	8	2	0,9	0,9	18	0,6077	0,5496	0,4143	1,2374
5	7	2	0,9	0,9	12	0,5874	0,5468	0,4004	1,1985
6	6	2	0,9	0,9	5	0,8855	0,9748	0,7266	1,09
7	5	2	0,9	0,9	4	1,0347	1,074	0,898	1,05
8	7	2	0,9	0,8	21	0,3769	0,2583	0,2694	1,2016
9	7	2	0,8	0,8	15	0,393	0,3289	0,3056	1,2019
10	6	2	0,9	0,8	12	0,4782	0,417	0,3569	1,1827
11	6	2	0,8	0,8	14	0,3603	0,3175	0,3069	1,1457
12	5	2	0,9	0,8	8	0,5264	0,4714	0,4109	1,1278
13	5	2	0,8	0,9	4	0,5711	0,6044	0,5165	0,988
14	5	2	0,8	0,8	8	0,499	0,4694	0,4227	1,125
15	5	2	0,8	0,7	4	0,67	0,7777	0,6544	1,025
16	5	2	0,7	0,7	19	0,4153	0,3279	0,3035	1,1918
17	5	2	0,7	0,75	14	0,4534	0,4268	0,3465	1,2047
18	5	2	0,7	0,78	14	0,4143	0,3419	0,3107	1,1507
19	6	2	0,8	0,9	15	0,397	0,3643	0,3147	1,2066
20	6	2	0,9	0,85	12	0,5121	0,5068	0,411	1,1896
21	7	2	0,7	0,7	34	0,3675	0,1944	0,2304	1,2473
22	7	2	0,8	0,7	35	0,2993	0,1542	0,2121	1,2472
23	5	2	0,9	0,85	7	0,5281	0,5092	0,4606	1,1044
24	5	2	0,8	0,86	7	0,495	0,4829	0,4217	1,1006

ниях количества кластеров, особенно для коэффициента Nэ*Na.

Третий вид коэффициента «С» (количество строк матрицы соответствия) признан наиболее правильным — на графике нет ярко выраженной зависимости от количества кластеров.

Выбор параметров алгоритмов кластеризации

Каждый эксперимент алгоритмов кластеризации сравнивался с каждым типом экспертной классификации.

Для алгоритма Кохонена до начала процесса кластеризации можно задавать следующие параметры:

- число нейронов;
- радиус активации;
- норма обучения;
- множитель нормы обучения.

Для FCM-алгоритма до начала процесса кластеризации задаются параметры:

- требуемое количество кластеров;
- экспоненциальный вес;
- количество итераций;
- значение целевой функции, при котором процесс завершается[1].

В таблицах 1 и 2 приведены параметры для второй части плана экспериментов алгоритмов Кохонена и FCM со значениями целевой функции, полученными при сравнении с каждым типом экспертной классификации.

Как видно из таблицы 1, лучшие значения оценочной функции при сравнении с типом экспертной классификации «Классы документации» (количество классов 2) дают эксперименты Кохонена с наименьшим количеством получившихся кластеров (4 и 5). Для остальных типов экспертных классификаций наиболее хорошие результаты единодушно получены в экспериментах 3 и 22. Таким образом, можно сделать вывод о наиболее подходящих параметрах кластеризации с помощью алгоритма Кохонена.

1 Данная целевая функция является принадлежностью FCM-алгоритма и не имеет отношения к целевой функции оценки процесса кластеризации.

2 В скобках указано количество классов, полученное в процессе экспертной классификации данного типа.

Таблица 2

Параметры и значения целевой функции для FCM-алгоритма второй части плана экспериментов[3]

№ эксп.	Кол-во кластеров	Эксп. вес	Значения оценочной функции для типов экспертной классификации			
			виды документации	изделия	разделы	классы документации
1	10	1,3	0,3482	0,3228	0,3309	1,0991
2	10	1,4	0,3618	0,3351	0,3569	1,0156
3	10	1,5	0,383	0,3409	0,3519	1,095
4	10	1,6	0,3847	0,3657	0,3317	1,124
5	13	1,3	0,353	0,2958	0,3129	1,1146
6	3	1,3	0,6261	0,742	0,6546	0,9162
7	3	1,4	0,5511	0,582	0,5436	0,8812
8	3	1,5	0,5194	0,5517	0,5292	0,8625
9	3	1,6	0,5194	0,5485	0,5212	0,8625
10	13	1,4	0,3524	0,3058	0,3193	1,0775
11	13	1,3	0,3557	0,316	0,3215	1,1338
12	10	1,3	0,3551	0,3294	0,3366	1,0775
13	10	1,3	0,3562	0,3278	0,3327	1,0991
14	10	1,3	0,3575	0,331	0,3382	1,1145
15	31	1,3	0,2896	0,2043	0,245	1,2061
16	31	1,3	0,3006	0,1997	0,2409	1,2248
17	31	1,3	0,2978	0,206	0,2422	1,2076
18	13	1,3	0,3569	0,318	0,3222	1,1468
19	13	1,3	0,3523	0,3128	0,3188	1,1335

3 Количество итераций всегда принималось равным 100, значение целевой функции алгоритма — равным 0,001.

Для FCM-алгоритма получаем, что при сравнении с экспертной классификацией с малым количеством классов лучшие результаты дают параметры: количество кластеров — 3 и экспоненциальный вес — 1,5 - 1,6. Для экспертных классификаций с достаточно большим количеством классов наилучшими параметрами являются: количество кластеров — 31 и экспоненциальный вес — 1,3.

Заключение

Таким образом, разработано программное приложение, обеспечивающее автоматизацию процесса экспертной классификации, реализацию метода построения матриц соответствия и вычисления значений оценочной функции для сравнения результатов экспертной классификации и автоматических кластеризаций.

Проведены два этапа экспериментов, определен нормировочный коэффициент оценочной функции, определены оптимальные параметры алгоритмов кластеризации.

СПИСОК ЛИТЕРАТУРЫ

1. Радионова Ю.А. Метод построения оценочной функции, определяющей эффективность алгоритмов автоматической кластеризации // Автоматизация процессов управления. — 2009. — №1(15). — С. 23—28.

2. Радионова Ю.А., Тронин В.Г. Классификация технической документации на основе лексического анализа децимального номера // Автоматизация процессов управления. — 2008. — № 3(13). — С. 69—72.