**УДК 519.95**

**Н.А.ИГНАТЬЕВ, Ш.Ю.НУРЖОНОВ (НУУз)**

**СЕЛЕКЦИЯ ОБУЧАЮЩИХ ВЫБОРОК НА ОСНОВЕ СВОЙСТВА СВЯЗАННОСТИ ОБЪЕКТОВ**

*Прецедент бўйича англаш алгоритмларининг умумлаштирувчи қобилиятини ошириш масаласи қаралади. Шовқин объектларни аниқлаш мезони ва танланмани минимал эталон - объектлар билан қоплаш учун кластерли таҳлилдан фойдаланиш таклиф қилинади. Кластерли таҳлил танланмани ўзаро кесишмайдиган гуруҳларга боғланганлик муносабати бўйича ажратиш синфларнинг қобиқ - чегаравий қисм тўплам объектлари орқали амалга оширилади. Машҳур FRiS-STOLP алгоритми билан умумлаштирувчи қобилияти бўйича солиштириш таҳлили қилинади.Ўргатувчи танланма селекцияси реал вақт режимида технологик жараёнларни бошқариш масалаларида талаб қилинади.*

***Таянч сўзлар:*** *ўргатувчи танланма, умумлаштирувчи қобилият, шовқин объектлар, вазнланган метрика.*

*Рассматривается проблема повышение обобщающей способности алгоритмов распознавания по прецедентам. Для решения предлагается использовать критерии удаления шумовых объектов и кластерный анализ выборки для минимального покрытие её объектами – эталонами. Кластерный анализ по результатам разбиения выборки на непересекающиеся группы по отношению связанности реализуется через оболочки-подмножества граничных объектов классов. Приводится сравнительный анализ обобщающей способности с известным алгоритмом FRiS-STOLP. Селекция обучающих выборок востребована в задачах управления технологическими процессами в реальном режиме времени.*

***Ключевые слова:*** *обучающая выборка, обобщающая способность, шумовые объекты, взвешенная метрика.*

*The problem of improving the generalization ability of pattern recognition algorithms by precedents is considered.* *For solution is proposed to use criteria removing noisy objects and cluster analysis of the sample for a minimum covering of its objects - prototypes.* *Cluster analysis by results of the sample partition into nonintersecting groups by relatedness is implemented through shell-subsets of the boundary class objects.* *We provide a comparative analysis by the generalization ability with known algorithm FRiS-STOLP.* *Selection of training samples is demanded by in problems process control in real time.*

***Key words:*** *training set, generalization ability, noisy objects, weighted metrics.*

Вопрос о селекции объектов обучающих выборок и влиянии её на обобщающую способность алгоритмов распознавания рассматривается как одна из актуальных задач в теории обучающих систем. Перечень подходов для решения проблемы обобщающей способности алгоритмов можно найти в [1]. Проблема цензурирования (селекции) путём отбора подмножества объектов обучающих выборок и наборов информативных признаков обсуждается в [2, 3].

Как правило, объекты обучения не являются равноценными. Среди них могут находиться эталоны - типичные представители классов. Отбор «столпов» (объектов-эталонов) с использованием функции конкурентного сходства (FRiS-функции) реализован в алгоритме STOLP [4]. Процесс отбора сводится к задаче локально-оптимального покрытия обучающей выборки. Поиск экстремума задачи о минимальном покрытии связан с перебором всевозможных вариантов. Количество столпов классов, гарантирующих отсутствие переобучения алгоритмов распознавания рассматривается в [2].

В работе для исследования обобщающей способности алгоритмов предлагается проводит кластерный анализ объектов классов на основе свойства связанности. Кластерному анализу предшествует процедура выделения шумовых объектов. Анализ структуры обучающей выборки по свойству связанности производится через подмножество граничных по заданной метрике объектов (оболочке) классов. Свойство связанности гарантирует единственность решения на обучающей выборке, при котором число групп и их состав остаётся неизменным.

Для селекции обучающих выборок в работе предложено:

* + два критерия для обнаружения шумовых объектов;
  + разбиение объектов классов на непересекающиеся группы для выбора объектов-эталонов покрытия.

**Постановка задачи**

Рассматривается множество допустимых объектов , разделённое на непересекающихся классов . Каждый объект описывается n разнотипными признаками ,  из которых измеряются в интервальной,  в номинальной шкалах. В качестве меры близости между объектами определена метрика . На множестве пар объектов вводится отношение связанности Объекты связаны отношением если и истинно.

Множество образует группу из связанных отношением объектов, если для любых существует путь (цепочка) и Требуется:

* для объектов каждого класса выделить подмножество ближайших (шумовых) по метрике объектов ;
* произвести разбиение множества () на непересекающиеся группы из связанных отношением объектов;
* из выделить объекты – эталоны, позволяющие корректно (без ошибок) распознавать объекты .

**Выделение шумовых объектов**

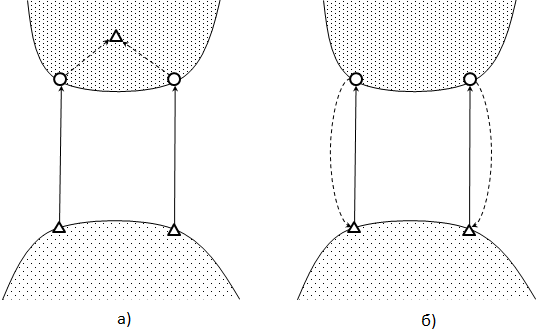
Пусть и . Объект включается в и рассматривается как шумовой если выполняется условие

где

. Альтернативным к (1) условием является

где Доля объектов включаемых в окрестность после удаления согласно (2) должна быть больше чем до его удаления.

Графическая иллюстрация конфигурации связей между каждым объектом класса и ближайшим объектом из противоположного класса показана на рис.1.



***Рис.1.******Отношения “объект-ближайший объект из противоположного***

***класса”: а) при наличии шумового объекта; б) при удалении шумового объекта.***

Целью селекции объектов обучающих выборки является повышение обобщающей способности алгоритмов. Считается, что обобщающая способность распознающего алгоритма повышается, если дать ему возможность ошибаться на некоторых определяемых объектах выборки. В качестве таковых рассматривается объекты из .

**Группировка объектов по отношению связанности**

Целью группировки объектов является:

* получить знание о топологии признакого пространства;
* выделить объекты – эталоны покрытия обучащей выборки.

Для поиска объектов - эталонов классов требуется определить:

* минимальное число групп из связанных объектов по каждому классу ;
* минимальное покрытие множества объектами–эталонами для алгоритмов распознавания по прецедентам.

Теоретические основы описываемого ниже алгоритма выделения объектов-эталонов изложены в [5]. Минимальное число групп связанных объектов классов определяется на основе предобработки данных. Предобработка данных заключается в:

* выделении оболочки – подмножества граничных объектов классов по заданной метрике [5];
* описании объектов каждого класса по своей системе бинарных признаков.

Для выделения оболочки классов для каждого построим упорядоченную по последовательность

Пусть () ближайший к объект из (3) не входящий в класс . Обозначим через окрестность радиуса с центром в , включающую все объекты, для которых . В всегда существует непустое подмножество объектов

По (4) принадлежность объектов к оболочке классов определяется как .

Множество объектов оболочки из обозначим как . Значение однозначно определяет вхождение всех объектов класса в одну область. При преобразуем описание каждого объекта в , где

Пусть по (5) получено описание объектов класса в новом (бинарном) признаковом пространстве, , – число непересекающихся между собой групп объектов, , – соответственно операции дизъюнкции и конъюнкции по бинарным признакам объектов . Пошаговое выполнение алгоритма разбиения объектов на непересекающиеся группы таково.

1 шаг: ;

2 шаг: Выделить объект , , , ;

3 шаг: **Выполнять** Выбор and , , ,

**пока** ;

4 шаг: Если , то идти 2;

5 шаг: Конец.

**Покрытие обучающей выборки объектами – эталонами**

Обозначим через расстояние от объекта до ближайшего объекта из противоположного к класса (), через – минимальное число групп из связанных объектов в . Для поиска минимального покрытия объектами–эталонами обучающей выборки упорядочим объекты каждой группы , , по множеству значений . В качестве меры близости между , и произвольным допустимым объектом используется взвешенное расстояние . Решение о принадлежности к одному из классов принимается по правилу: если

Согласно принципа *последовательное исключение*, используемого в процессе поиска покрытия, выборка делится на два подмножества: множество эталонов и контрольное множество , . В начале процесса , . Упорядочение по значениям отступа , используется для определения кандидата на удаление из числа объектов–эталонов по группе . Идея отбора заключается в поиске минимального числа эталонов, при котором алгоритм распознавания по (6) остаётся корректным (без ошибок распознающим объекты) на .

Будем считать, что нумерация групп из связанных между собой объектов отражает порядок и по группе , не производился отбор эталонных объектов. Кандидаты на удаление из последовательно выбираются начиная с с минимальным значением . Если включение нарушает корректность решающего правила (6), то возвращается в множество .

**Вычислительный эксперимент**

Вычислителный эксперимент проводился на выборке данных IONOSHPERE из [7]. Выборка представлено 351 объектами, каждый из которых описывается 33 количественными признаками.

В табл.1, табл.2, табл.3 представлены результаты использования метрики Чебышева как с учётом удаления шумовых объектов по критерию (1), (2) так без такого учёта.

*Таблица 1.* ***Удаление шумовых объектов по (1)***

|  |  |  |  |
| --- | --- | --- | --- |
| Значение | Количество | | |
| шумовых объктов | Групп | эталонов |
| 0 | 27 | 10 | 21 |
| 1 | 20 | 8 | 20 |
| 2 | 19 | 8 | 20 |
| 3 | 18 | 8 | 20 |
| 4 | 17 | 8 | 20 |
| 5 | 13 | 9 | 19 |

*Таблица 2.* ***Удаление шумовых объектов по (2)***

|  |  |  |  |
| --- | --- | --- | --- |
| Значение | Количество | | |
| шумовых объктов | Групп | эталонов |
| 0.1 | 34 | 7 | 21 |
| 0.5 | 32 | 7 | 20 |
| 0.6 | 32 | 7 | 20 |
| 0.8 | 31 | 7 | 18 |
| 0.9 | 30 | 7 | 18 |
| 0.99 | 23 | 6 | 21 |

*Таблица 3.* ***Результаты без учёта шумовых объектов***

|  |  |
| --- | --- |
| Количество | |
| групп | Эталонов |
| 13 | 27 |

Анализ содержимого табл.1, табл.2, показывает близость результатов, полученных при использовании критериев (1), (2) и значительное различие их по количеству групп и эталонов из табл.3, полученных без селекции обучающей выборки.

Обобщающая способность алгоритмов по объектам-эталонам, полученным без удаления и с удалением шумовых объектов (обозначаемых как QOBIQ1 и QOBIQ2) и по алгоритму FRISSTOLP, представлена в табл.4. Для расчёта использовался метод скользящего контроля при соотношении обучение:контроль=9:1. Выбор объекта в качестве эталона проводилась по критерию эффективности

,

где , – характеристики соответственно обороноспособности и толерантности, параметр

***Таблица 4. Обобщающая способность алгоритмов распознавания***

|  |  |  |  |
| --- | --- | --- | --- |
| Алгоритм | Базовая метрика | | |
| Хэмминга | Чебышева | Евклида |
| QOBIQ1 | 68.28 | 68.57 | 64.85 |
| QOBIQ2 | 74.86 | 78.57 | 74 |
| FRISSTOLP () | 71.43 | 62.29 | 66.57 |
| FRISSTOLP () | 75.14 | 71 | 65 |

Из табл.4 следует, что наибольшая обобщающая способность (78.57%) достигается по базовой метрике Чебышева при удаление шумовых объектов.

**Заключение.**

При управлении технологическими процессами в реальном режиме времени требуется высокая точность классификации их состояний. Для принятия решения требованиям высокой точности и минимальной сложности при распознавании по прецедентам отвечает описанный в работе алгоритм метода селекции обучающих выборок.

**Список литературы:**

1. Воронцов К.В. Обзор современных методов по проблеме качества обучения алгоритмов //Таврический вестник информатики и математики, 2004. №1. С. 5-25.
2. Загоруйко Н.Г., Кутненко О.А. Цензурирование обучающей выборки //Вестник Томского государственного университета. 2013. №1(22). С.66-73.
3. Загоруйко Н.Г., Кутненко О.А., Зырянов А.О., Леванов Д.А. Обучение распознаванию образов без переобучения // Машинное обучение и анализ данных. 2014. Т. 1 №7. С.891-901.
4. Загоруйко Н. Г. Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999.
5. Игнатьев Н.А. Обобщенные оценки и локальные метрики объектов в интеллектуальном анализе данных. –М.: ”Университет”, 2014. – 72с.
6. Asuncion A., Newman D. J. UCI Machine Learning Repository // University of California, Irvine. 2007. [www.ics.uci.edu/~mlearn/MLRepository.html](http://www.ics.uci.edu/~mlearn/MLRepository.html)
7. http://www.machinelearning.ru/wiki/index.php?title=Алгоритм\_FRiS-СТОЛП

*Игнатьев Николай Александрович – физика-математика фанлари доктори, профессор,*

*Ўзбекистон миллий университети «Программалаш ва тармоқ технологиялари» кафедраси профессори,*

*Тел.:(+99871)246-16-27 (р.). E-mail: n\_ignatev@rambler.ru.*

*Нуржонов Шерзод Юлдашович –* *Ўзбекистон миллий университети «Программалаш*

*ва тармоқ технологиялари» кафедраси стажер-тадқиқотчи-изланувчиси*

*Тел.: (+99890)974-24-27 (м.). E-mail: sherzod.n@mail.ru.*

*Игнатьев Николай Александрович – доктор физико-математических наук, профессор,*

*профессор кафедры «Программные и сетевые технологии» Национального университета Узбекистана,*

*Тел.:(+99871)246-16-27 (р.). E-mail: n\_ignatev@rambler.ru.*

*Нуржонов Шерзод Юлдашович –* *стажер-исследователь-соискатель кафедры «Программные*

*и сетевые технологии» Национального университета Узбекистана*

*Тел.: (+99890)974-24-27 (м.). E-mail: sherzod.n@mail.ru.*

*Ignat’ev Nikolay Alexandrovich - doctor of sciences, professor, department of "Software and*

*Network Technology" of the National university of Uzbekistan,*

*Tel.:(+99871)246-16-27 (w.). E-mail: n\_ignatev@rambler.ru.*

*Nurjonov Sherzod Yuldashovich -* *Senior Researcher, department of "Software and*

*Network Technology" of the National university of Uzbekistan,*

*Tel.: (+99890)974-24-27 (м.). E-mail: sherzod.n@mail.ru.*