УДК 004.093

РАЗРАБОТКА АЛГОРИТМ ФОРМИРОВАНИЕ СИСТЕМЫ ОПОРНЫХ МНОЖЕСТВ ПРИЗНАКОВ, ОБЕСПЕЧИВАЮЩИХ КАЧЕСТВО И НАДЕЖНОСТЬ РАСПОЗНАВАНИЯ

Бекмуратов Д.К.
старший преподаватель,
Самаркандский филиал Ташкентского университета информационных технологий,
тел.: +(99893) 727-99-77, e-mail: bekmurodov_d@mail.ru

В данной статье приводится методика нахождения предельнодопустимой размерности пространства комбинации признаков. Предложены принципы формирования системы опорных множеств с учетом предельнодопустимой размерности комбинации признаков относительно конкретного образа. Разработан алгоритм и программное обеспечение, обеспечивающее требуемые качества и надежность распознавания объектов.

Ключевые слова: объект, признак, класс, эталонная выборка, контрольная выборка, объем эталонной выборки, системы опорных множеств, качество и надежность распознавания, решающее правило.

DEVELOPMENT ALGORITHM FORMATION OF THE SYSTEM OF SUPPORT SETS OF SIGNS, PROVIDING QUALITY AND RELIABILITY OF RECOGNITION

Bekmuratov D.Q.

In this article, the technique of finding the maximumallowable dimensions of the feature space. The principles of formation of the system of support sets based extremely allowable dimension combinations of features on a specific image. The algorithm, which provides the required object recognition for quality and objects.

Keywords: object, attribute, class, reference sample, control sample, the scope of the reference sample, the system of reference sets, the quality and reliability of recognition, the decision rule.

OBYEKTLARNI ANGLASHDA SIFAT VA ISHONCHLILIKNI TAMINLAB BERUVCHI TAYANCH BELGILAR TO'PLAMI TIZIMINI SHAKLLANTIRISH ALGORITMINI ISHLAB CHIQISH

Bekmuratov D.Q.

Ushbu maqolada belgilar kombinatsiyasining maksimal qaralgan o'lchamini topish uchun texnik uslublar berilgan. Timsollarga nisbatan o'ziga xos xususiyatlar kombinatsiyasini maksimal qaralgan o'lchamga moslashtiruvchi va tayanch belgilar tizimini shakllantirish printsiplari taklif qilingan. Obyektlarni aniqlashda sifat va ishonchliligini ta'minlaydigan tayanch belgilar tizimini shakllantirish algoritmi va dasturiy ta'minoti ishlab chiqilgan.

Kalit so'zlar: obyekt, beldilar, sinflar, etalon tanlov, nazorat tanlov, etalon tanlov chegarasi, mos to'plamlar tizimi, sifatli va ishonchli anglovchi tizim, hal qilivchi qoida.

1. Введение

Существуют различные методы установления критериев информативности, позволяющие определить информативные признаки или сочетаний признаков. Однозначно отдать предпочтение одному из них невозможно. Методы, лучше работающие для одного из классов задач, хуже приемлемы для другого класса, вместе с тем, во многих методах при установлении критерия информативности признаков или сочетаний признаков не учитывается объем обучающей выборки. Исходя из таких соображений, выбор определенного метода с известной долей интуиции должен производиться в зависимости от реальной задачи и конкретных практических возможностей. Более того, при установлении критериев информативности признаков или сочетаний признаков нельзя не учитывать такие важные параметры, как качество и надежность распознавания.

Для обеспечения высокого качества распознавания нужно знать признаки объектов, в которых предстоит работать, распознающей системе, а обучающую последовательность следует выбирать так, чтобы признаки объектов в полной мере отразились в той последовательности [1-3].

Таким образом, распознающая система должна обладать не только заданным качеством, но и надежностью его достижения. Задача построения распознающей системы сводится к тому, чтобы в процессе обучения по обучающей последовательности было достигнуто определенное

качество, достижение которого обеспечивалось бы с надежностью, не ниже заданной.

С целью достижения определенного качества распознающих систем производится предварительное сокращение размерности обучающей выборки. Из первоначального алфавита признаков выбираются лишь несколько наиболее информативных признаков или сочетаний признаков. Сокращение размерности пространства признаков представляется полезным в двух аспектах: во-первых, снижается объем вычислений, а во-вторых, с удалением из эталонной выборки несущественных признаков повышается надежность распознавания [1-3]. Одновременно, за счет снижения объема эталонной выборки уменьшается объем, что приводит зачастую к снижению надежности распознавания в целом. Иначе говоря, объем эталонной выборки должен находиться в разумных пределах.

В связи с изложенным обстоятельством создание для каждого из наиболее распространенных критериев информативности "своего" метода оптимизации, который учитывал бы математические признаки конкретного критерия, приобретает весьма актуальный и своевременный смысл.

В работе [1-3] получены теоретические результаты, заставляющие весьма осторожно относиться к большинству известных методов распознавания, не уделяющих специального внимания формированию признакового пространства, в котором в процессе обучения строятся решающие правила. Эти результаты основаны на соотношениях, связывающие такие параметры, как количество объектов и признаков обучающей выборки, качество и надежность распознавания [1, 3]. В этих работах показано, что чем проще решающее правило и чем ниже размерность признакого пространства, тем меньше вероятность ошибочных ответов, возникающих при эксплуатации распознающей системы.

Эти выводы можно рассматривать как обоснование главной цели настоящей статьи.

2. Постановка задачи

В [3] получен теоретический результат, смысл которого состоит в том, что если на эталонной выборке из N решающих правил выбирается одно, которое безошибочно разделяет эталонную выборку длины m, то с вероятностью $(1-\eta)$ можно утверждать, что вероятность ошибки при распознавании объектов с помощью этого правила составит величину, меньшую ε, где

$$\varepsilon = \frac{\ln N - \ln \eta}{m} . \qquad (1)$$

Из (1) следует, что чем проще решающее правило и чем ниже размерность признакого пространства, тем меньше вероятность ошибочных ответов, возникающих при эксплуатации распознающей системы.

В работах [4, 5] показано, что системы опорных множеств Ω_A строятся следующим образом:

а) $\Omega_A = \{\Omega : |\Omega| = k\} = C_n^k$;

b) $\Omega_A = \{\Omega\}, \Omega \subseteq \{1,2,...,n\} = C_n^1 + C_n^2 + ... + C_n^n = 2^n$

В случае а) значение k находится из решения задачи обучения (оптимизации модели) или задается экспертом.

В случае б) способ выбора системы опорных множеств, как всевозможных подсистем *{1,2,...,n},* является «усреднением» первого и не требует нахождения подходящего значения параметра *k.*

В данной статьи, в отличие от [4, 5], при построении системы опорных множеств Ω_A используются не все комбинации, а только $C_n^{n_0}$ $(n_0 = 1; n_0 = 2; ...; n_0 = k; (k < n))$ комбинации признаков. Здесь $n_0 (n_0 < n)$ заранее определяется с учетом требуемых значений качества и надежности распознавания при заданном объеме (число признаков и объектов) эталонной выборки, т.е. $n_0 = f(\varepsilon, \eta, n, m)$, где ε - вероятность ошибки, η - надежность распознавания, n - количество признаков, m - количество объектов. Это приводит к резкому сокращению системы опорных множеств Ω_A и позволяет при распознавании объектов использовать комбинации признаков входящих в $\Omega_A = C_n^{n_0}$.

Таким образом, для успешного решения любой задачи распознавания в алгоритмах вычисления оценок [4, 5] необходимо стремиться к минимизации системы опорных множеств Ω_A, снижению (в пределах n_0) размерности пространства комбинаций признаков и упрощению класса решающих правил (в пределах Ω_A). Поэтому из эталонной выборки необходимо формировать такие системы опорных множеств Ω_A, которые обеспечивают требуемое качество и надежность распознавания.

Пусть задана эталонная выборка объектов $S_{j1}, S_{j2}, ..., S_{jm_j} \in K_j, j = \overline{1,l}$, где каждый объект S_j является n - мерным вектором числовых признаков, т.е. $S_j = (x_{j1}, ..., x_{jn}), j = \overline{1,m}$ ($m = m_1 + m_2 + ... + m_l$).

Пусть для образов $K = K_1, ..., K_l$ выполняется $K_i \cap K_j = \phi, \forall i \neq \forall j$. Обозначим через T_{nml} - эталонную выборку, $T^*_{nm_1}$ - контрольную выборку, где n -количество признаков, m - количество объектов, l - количество классов, m_1 - количество объектов контрольной выборки.

Требуется, используя эталонную выборку T_{nml}, найти предельно-допустимую размерность пространства комбинаций признаков n_0, построить систему опорных множеств Ω_A из комбинаций признаков $C_n^{n_0}$ и определить на этом множестве решающие правила, обеспечивающие требуемые качество и надежность распознавания.

3. Формирование системы опорных множеств

В [4, 5] системы опорных множеств из комбинаций признаков в первом случае определяется как $\Omega_A = C_n^k$ и во втором случае как $\Omega_A = C_n^1 + C_n^2 + ... + C_n^n = 2^n$. Если учесть, что каждый объект $S_i \in T^*_{nm_1}$ сопоставляется с каждым объектом $S_j \in T_{nml}$ с помощью $C_n^i (i = \overline{1,k}$ либо $i = \overline{1,n})$ на основе правила

$$d(S_i, S_j) = \begin{cases} 1, если \ \ \tilde{\omega}\mathrm{S_i} = \tilde{\omega}S_j \\ 0, \text{ если } \ \tilde{\omega}S_i \neq \tilde{\omega}S_j \end{cases}, \qquad (2)$$

тогда результаты сопоставления будут соответствовать одному из вариантов 2^i, где $\tilde{\omega}\mathrm{S_i}$ и $\tilde{\omega}\mathrm{S_j}$ называются ω-частью объектов $\mathrm{S_i}$ и $\mathrm{S_j}$ соответственно. Следовательно, из [4, 5] следует, что если учесть в первом случае $\Omega_A = C_n^k$ и 2^i, во втором случае $\Omega_A = C_n^1 + C_n^2 + ... + C_n^n = 2^n$ и 2^i, тогда число всевозможных подмножеств $\{1,2,...,n\}$ длины i и результаты вариантов сопостовления объектов $\mathrm{S_i}$ и $\mathrm{S_j}$ будут $N = 2^i C_n^i,\ i = \overline{1,k}$ и $N = 2^i C_n^i,\ i = \overline{1,n}$ соответственно.

Если в эталонной выборке количество объектов и признаков задаётся слишком много, то в компьютере для сопостовления объектов на наборах $N = 2^i C_n^i,\ i = \overline{1,n}$ требуется много времени. Поэтому в формуле $N = 2^i C_n^i,\ i = \overline{1,n}$ из $i = 1, i = 2, ..., i = n_0, ..., i = n$ нужно найти конкретное значение i (например $i = n_0$), и необходимо получить 2^{n_0} различных результатов при сопоставлении объектов в системе опорных множеств $\Omega_A = C_n^{n_0}$ эквивалентных результатам, получаемым с помощью $N = 2^i C_n^i,\ i = \overline{1,n}$. Для достижения этой цели требуется найти конкретное значение n_0 [7, 9].

Если из эталонной выборки T_{nml} определена система опорных множеств $\Omega_A = C_n^{n_0}$ и всевозможные варианты результатов 2^{n_0} сопоставления объектов $\mathrm{S_i} \in T^*_{nm_1}$ с объектами $\mathrm{S_j} \in T_{nml}$, то число всевозможных решающих правил составит величину, меньшую N, где

$$N = 2^{n_0} C_n^{n_0}. \qquad (3)$$

Предположим, что до обучения заранее заданы требуемые значения вероятности ошибки ε и надежность распознавания η, а также количество признаков n и объектов m. Тогда из (1) можно получить функциональную зависимость

$$ln\, N = \varepsilon m + ln\, \eta. \qquad (4)$$

Для того чтобы найти n_0 логарифмируем (3)

$$\ln N = \ln 2^{n_0} + \ln C_n^{n_0}. \qquad (5)$$

Используя $C_m^n \le \frac{m^n}{2^n}$ и учитывая (6) из соотношения (5) получим

$$\ln N = n_0 \ln n\eta. \qquad (6)$$

Если полученное значение (6) подставить в соотношение (4), то можно получить конкретную функциональную зависимость для n_0

$$n_0 = \frac{\varepsilon m + \ln \eta}{\ln n}. \qquad (7)$$

Найденное по этой зависимости значение n_0 гарантирует заданную вероятность ошибки ε с надежностью $(1-\eta)$ при фиксированных m, n, η.

Если зафиксировать η, m, n, ε, то из соотношения (7) можно найти предельные значения размерности n_0 пространства комбинаций признаков, удовлетворяющие заданной вероятности ошибки ε при распознавании новых объектов (табл. 1) [6].

Таблица 1

$\eta = 0.95$, $m = 100$, $n = 50$.					
ε	0,04	0,07	0,12	0,15	0,19
n_0	1	2	3	4	5

Анализ приведенных в таблице 1 данных показывает, что увеличение вероятности ошибки ε при распознавании новых объектов, приводит к увеличению размерности n_0 пространства комбинаций признаков.

Если зафиксировать η, n_0, ε, n, то из соотношения (7) можно найти требуемое количество объектов $m = \frac{n_0 \ln n + \ln \eta}{\varepsilon}$, удовлетворяющих заданной вероятности ошибки ε при распознавании новых объектов (табл. 2).

Таблица 2

$\eta = 0.95$, $n = 20$.					
n	1	2	3	4	5
ε	0,02	0,03	0,04	0,05	0,06
m	294	297	298	298	299

Анализ приведенных в таблице 2 данных показывает, что с увеличением вероятности ошибки ε при распознавании новых объектов и размерности n_0 пространства комбинаций признаков, приводит к увеличению требуемых количеств объектов m.

4. Алгоритм решения задачи

Рассмотрим алгоритм, позволяющий формировать систему опорных множеств $\Omega_A = C_n^{n_0}$, определяющих в этом множестве решающие правила

$F(S_j),(S_j \in T^*_{nm_1})$, которые обеспечивают требуемое качество и надежность распознавания.

Алгоритмом сначала из эталонной выборки T_{nml} формируется система опорных множеств $\Omega_A = C_n^{n_0}$, далее в этом множестве каждый объект $S_j \in T^*_{nm_1}, j = \overline{1,m_1}$ сопоставляется с объектами $S_1,...,S_{m_j} \in K_j, j = \overline{1,l}$ и в результате суммарного голосования (суммарной степени близости распознаваемого объекта S_i к классу K_j) каждый объект $S_j \in T^*_{nm_1}, j = \overline{1,m_1}$ классифицируется, одному из заранее заданных классов $K_j \in T_{nml}, j = \overline{1,l}$. При этом в результате классификации новых объектов $S_j \in T^*_{nm_1}, j = \overline{1,m_1}$ обеспечиваются требуемые ε и η с учетом заданных n, m.

Алгоритм включает в себя следующие основные этапы:

1. В оперативную память заносятся объекты $S_1, S_2, ..., S_m$, их признаки $x_1, x_2, ..., x_n$ и классы $K_1, K_2, ..., K_l$ в виде эталонной выборки T_{nml}.

2. В оперативную память заносятся объекты $S_1, S_2, ..., S_{m_1}$ и их признаки $x_1, x_2, ..., x_n$ в виде контрольной выборки $T^*_{nm_1}$.

3. Определяется предельное значение допустимой размерности n_0 в виде (7), с учетом заданных значениях ε и η, а также при фиксированных n и m.

4.Формируются Ω_A из заданных n признаков по n_0, т.е. $\Omega_A = C_n^{n_0}$.

5. $i = 1$. Из оперативной памяти отбирается объект $S_i \in T^*_{nm_1}$.

6. $j = 1$. Отбирается объект $S_j \in T_{nml}$.

7. Сопоставляется объект $S_i \in T^*_{nm_1}$ с объектом $S_j \in T_{nml}$ по правилу (2).

8. Ставится объекту $S_j \in T_{nml}$ 0 (ноль) либо 1 (единица) в зависимости от результатов схожести, который производиться в 7-шаге алгоритма.

9. $j = j + 1$. Если $j \le m$, то алгоритм переходит к шагу 6, в противном случае к шагу 10.

10. Формируется новая таблица T_{Ω_A}, полученных по системе опорных множеств $\Omega_A = C_n^{n_0}$.

11. $p = 1$. Выделяется класс K_p и относительно этого класса вычисляется сумма голосов $\Gamma_\Sigma(K_p)$.

12. $\Gamma_\Sigma(K_p) = 0$.

13. $j = 1$. Выделяется объект $S_j \in K_p$.

14. $\Gamma_{\Omega_A}(S_j) = 0$.

15. $k = 1$.

16. Проверяется условие:

а) Если $d(\tilde{\omega}S_{ik}, \tilde{\omega}S_{jk}) = 1$, то, и алгоритм переходит к шагу 17.

б) Если $d(\tilde{\omega}S_{ik}, \tilde{\omega}S_{jk}) = 0$, то алгоритм переходит к шагу 17.

17. $k = k + 1$. Если $k \le t (t = C_n^{n_0})$, то алгоритм переходит к шагу 15, в противном случае к шагу 18.

18. $j = j + 1$. Если $j \le m_i$, то алгоритм переходит к шагу 13, в противном случае к шагу 19.

19. $p = p + 1$. Если $p \le l$, то алгоритм переходит к шагу 11, в противном случае к шагу 20.

20. Вычисляются сумма голосов схожести распознаваемого объекта S_i к классу K_p:

$$\Gamma_\Sigma(S_i \in K_1) = \sum_{j=1}^{m_1}\sum_{k=1}^{t}\Gamma_{\Omega_A}(S_{jk})$$

$$\Gamma_\Sigma(S_i \in K_2) = \sum_{j=1}^{m_2}\sum_{k=1}^{t}\Gamma_{\Omega_A}(S_{jk})$$

.....

$$\Gamma_\Sigma(S_i \in K_l) = \sum_{j=1}^{m_l}\sum_{k=1}^{t}\Gamma_{\Omega_A}(S_{jk})$$

21. Для распознавания объекта $S_i \in T^*_{nm_1}$ используется решающее правило

$$F(S_i): \begin{cases} \mathrm{S_i} \in K_p, \textit{ если } \max\limits_{1 \le p \le l} \begin{Bmatrix} \Gamma_\Sigma(S_i \in K_1), \\ \Gamma_\Sigma(S_i \in K_2), .., \\ \Gamma_\Sigma(S_i \in K_l) \end{Bmatrix} \\ \mathrm{S_i} \notin K_1, \mathrm{S_i} \notin K_2, ..., \mathrm{S_i} \notin K_l \\ \textit{если } \Gamma_\Sigma(S_i \in K_1) = \\ \Gamma_\Sigma(S_i \in K_2) = \\ ... = \Gamma_\Sigma(S_i \in K_l) \end{cases}$$

22. $i = i + 1$. Если $i \le m_1$, то алгоритм переходит к шагу 5, в противном случае к шагу 23.

23. Вывод результатов, относящих объект $S_i \in T^*_{nm_1}$ по сумме голосований за классы, в один из классов $K_1, K_2, ..., K_l \in T_{nml}$, или указывающее для объекта $S_i \in T^*_{nm_1}$ отказ от распознавания.

5. Программное средство

Создан комплекс программ на основе разработанного алгоритма. Общий вид интерфейсного окна программы имеет следующий вид (рис. 1).

Модуль вычисления n_0, системы опорных множеств $\Omega_A = C_n^{n_0}$ и распознавания объектов по $\Omega_A = C_n^{n_0}$ представлен на рис. 2.

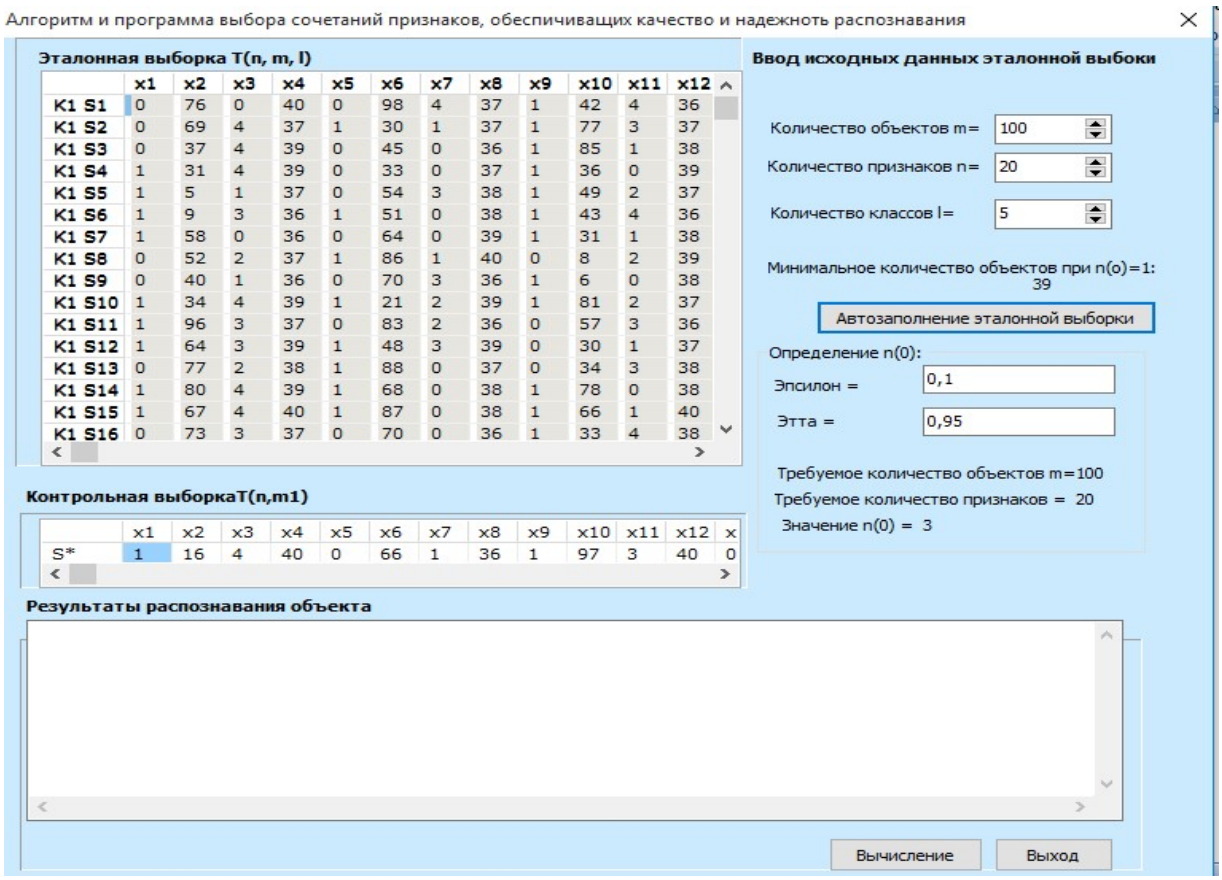

Рис. 1. Общий вид интерфейсного окна

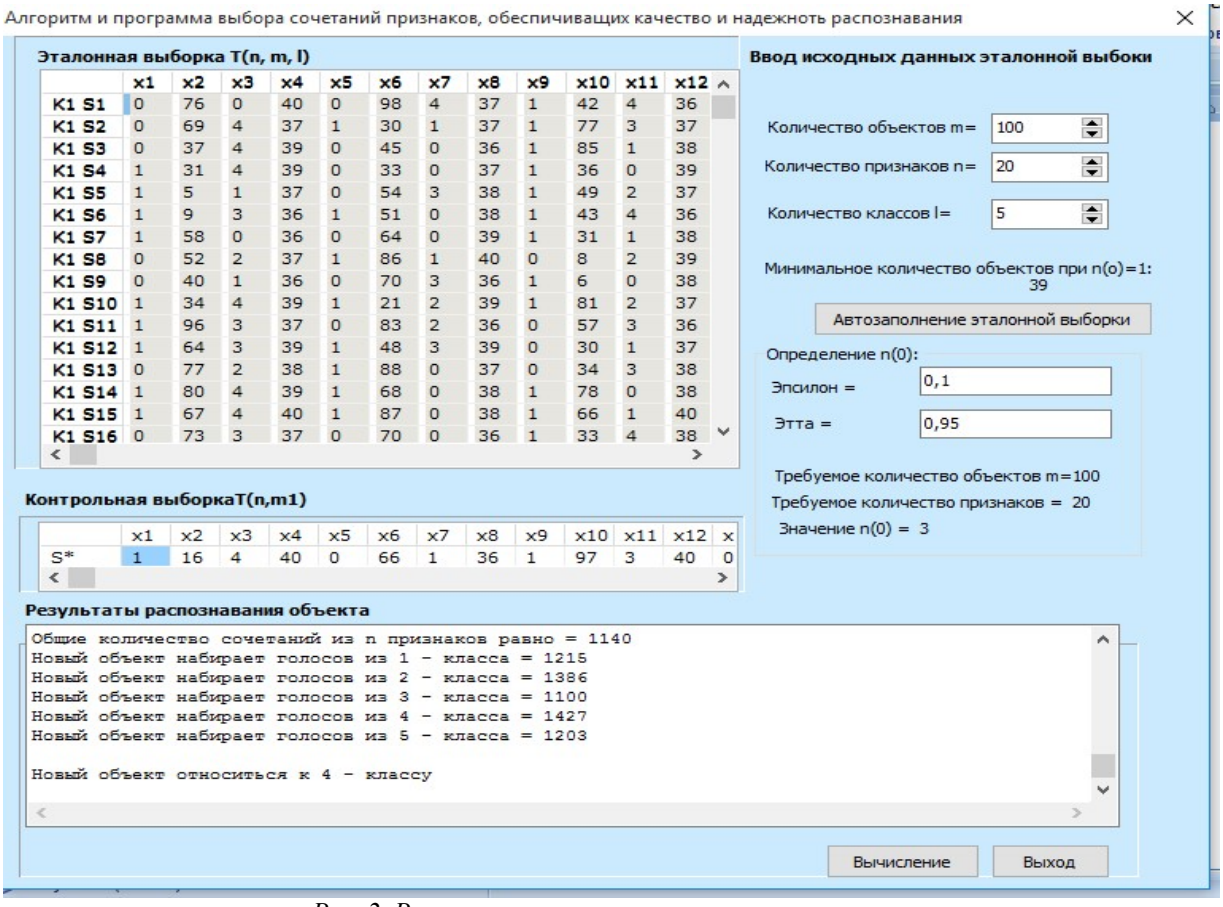

Рис. 2. Результаты вычислительного эксперимента

Проведено испытание по оценке работоспособности и эффективности предложенного алгоритма и программного комплекса применительно к распознаванию образов. Полученные результаты подтверждают то, что разработанный алгоритм и программный комплекс применим для решения практических задач распознавания объектов, касающихся медицинской, технической, археологической, гидрогеологической, сейсмологической, биологической и геологической сферы

6. Заключение и выводы

В отличие от алгоритмов, приведенных в [4, 5], в данном алгоритме:

- при заданных ε и η, а также при фиксированных n и m определяется n_0 в виде (7);

- формируется система опорных множеств из заданных n признаков по n_0, т.е. $\Omega = C_n^{n_0}$ $(n_0 \le n)$;

- вычисляется $\Omega_A = C_n^{n_0}$ сумма голосов схожести $\Gamma(S_j) \in T^*_{nm_1}, j = \overline{1, m_1}$ для каждого нового объекта $S_j \in T^*_{nm_1}, j = \overline{1, m_1}$;

- классифицируются объекты $\Gamma(S_j) \in T^*_{nm_1}$, $j = \overline{1, m_1}$ по сумме результатов голосования, одному из заданных классов $K_j \in T_{nml}, j = \overline{1, l}$.

- резко снижается объем вычислений на компьютере, так как

$C_n^{n_0} \le C_n^k$ $(n_0 < k)$ и $C_n^{n_0} \le C_n^1 + C_n^2 + ... + C_n^n$ $(n_0 < n)$.

Надо отметить, что в данном алгоритме сильно сокращаются системы опорных множеств Ω_A. Сокращение системы опорных множеств Ω_A представляется полезным в двух аспектах: во-первых, снижается объем вычислений, а во-вторых, с удалением из Ω_A лишних комбинаций признаков повышается надежность распознавания. Одновременно, за счет сокращения системы опорных множеств уменьшается множество, что приводит зачастую к снижению надежности распознавания в целом. Поэтому, в этом алгоритме, рассматривается их функциональная зависимость (5), чтобы качество и надежность распознавания, а также объем (количество признаков и объектов) эталонной выборки находился в требуемом интервале.

Литература

[1] *Бекмуродов К.А.*, *Васильев В.И.*, *Бекмуратов Д.К.* Нахождение предельно-допустимых значений размерности признаковых пространств из обучающей выборки // Современное состояние и перспективы развития информационных технологий / Академия Наук Республики Узбекистан, Институт математики и информационных технологий. Т. 2. – Ташкент, 2011. – С. 309-312.

[2] *Бекмуродов Қ.А.*, *Бекмуратов Д.Қ.* Последовательный выбор признаков, обладающих требуемой разделяющей силой // Научные перспективы XXI века. Достижения и перспективы нового столетия : по материалам XI - международной научно-практической конференции : ежемесячный научный журнал. Ч. 4. – Новосибирск, 2015. – №4(11). – С. 9-13.

[3] *Вапник В.Н.*, *Червоненкис А.Я.* Теория распознавания образов : статистические проблемы обучения. – М.: Наука, 1974. – 412 с.

[4] *Журавлев Ю.И.*, *Рязанов В.В.*, *Сенько О.В.* Распознавание. Математические методы. Программная система. Практические применения. – М.: ФАЗИС, 2005. –159 с.

[5] *Журавлев Ю.И.*, *Камилов М.М.*, *Туляганов Ш.Е.* Алгоритмы вычисления оценок и их применение. – Ташкент: ФАН, 1974. – 119 с.

[6] *Васильев В.И.* Распознающие системы. – Киев.: Наукова Думка, 1986. – 415 с.

[7] *Абдукаримов Р.Т.*, *Камилов М.М.*, *Кондратьев А.И.* Информационно-распознающие системы частичной прецедентности. – Т.: Фан, 1984. – 102 с.

[8] *Вапник В.Н.*, *Червоненкис А.Я.* Теория распознавания образов. – М.: Наука, 1974. – 415 с.

[9] *Горелик А.Л.*, *Скрипкин В.А.* Методы распознавания. – М.: Высшая школа, 2004. – 261 с.