УДК 004.855.5

ГИБРИДНАЯ ИНТЕЛЛЕКТУАЛЬНАЯ ТЕХНОЛОГИЯ КЛАССИФИКАЦИИ ДАННЫХ[1]

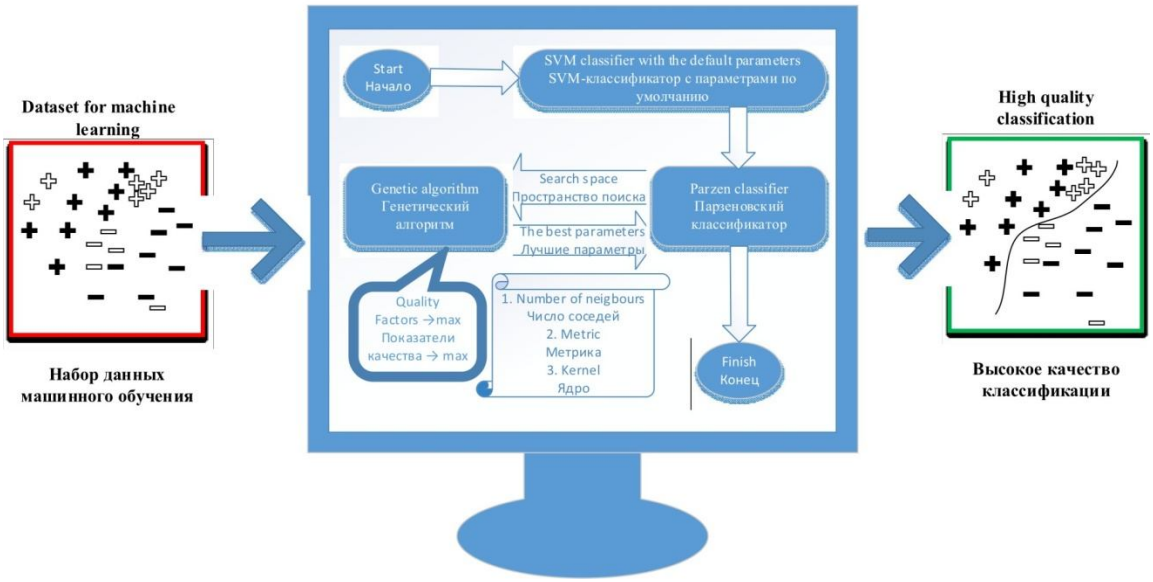
Статья поступила в редакцию 18.03.2018, в окончательном варианте 08.05.2018.

Демидова Лилия Анатольевна, Рязанский государственный радиотехнический университет, 390005, Российская Федерация, г. Рязань, ул. Гагарина, 59/1,
доктор технических наук, профессор, e-mail: liliya.demidova@rambler.ru
Егин Максим Михайлович, Рязанский государственный радиотехнический университет, 390005, Российская Федерация, г. Рязань, ул. Гагарина, 59/1,
студент, e-mail: eginmm@gmail.com

Рассматривается задача бинарной классификации наборов данных различной природы, представленных числовыми значениями характеристик. Эта задача решается с применением инструментария интеллектуального анализа данных. Целью работы является создание гибридной интеллектуальной технологии классификации данных (ГИТКД), основанной на совместном использовании SVM-алгоритма и метода окна Парзена. Классификатор на основе метода окна Парзена (КоМОП) позволяет улучшить точность классификации данных по сравнению с вариантом выполнения, использующим классификатор на основе SVM-алгоритма (КоSVM). КоМОП применяется к данным, которые могут быть классифицированы с использованием КоSVM, как верно, так и ошибочно. Эти данные находятся в экспериментально определяемых подобластях вблизи разделяющей классы гиперплоскости. При этом КоSVM используется со значениями параметров, заданными по умолчанию, а субоптимальные значения параметров КоМОП определяются с применением генетического алгоритма. Приведены результаты вычислительных экспериментов, проведенных на наборах данных из области медицинской диагностики, кредитного скоринга и обработки сигналов. Эти результаты подтверждают эффективность использования предлагаемой ГИТКД.

Ключевые слова: классификатор на основе SVM-алгоритма, опорный вектор, тип функции ядра, параметр функции ядра, параметр регуляризации, классификатор на основе метода окна Парзена, метрика расстояний, гибридная технология, генетический алгоритм, вычислительные эксперименты

Графическая аннотация (Graphical annotation)



HYBRID INTELLIGENT TECHNOLOGY OF DATA CLASSIFICATION

The article was received by editorial board on 18.03.2018, in the final version – 08.05.2018.

Demidova Liliya A., Ryazan State Radio Engineering University, 59/1 Gagarin St., Ryazan, 390005, Russian Federation,
Doc.Sci. (Engineering), Professor, e-mail: liliya.demidova@rambler.ru
Egin Maksim M., Ryazan State Radio Engineering University, 59/1 Gagarin St., Ryazan, 390005, Russian Federation,
Student, e-mail: eginmm@gmail.com

[1] Работа поддержана грантом РФФИ, номер заявки 17-29-02198.

The paper considers the problem of binary classification of the data sets of various nature, presented by numerical values of characteristics. This problem is solved with use of data mining tools. The aim of the paper is to create the hybrid intellectual technology for data classification (HITDC) based on the joint use of the SVM algorithm and the Parzen windows. The classifier, based on the Parzen windows (CbPWM), improves the accuracy of data classification performed using the classifier based on the SVM algorithm (CbSVM). The CbPWM applies to data that can be both correctly and erroneously classified using the CbSVM. These data are located in the experimentally defined subareas near the hyperplane separating the classes. The technology implies the use of default parameters values for CbSVM, while the suboptimal parameters values of the CbPWM are determined using the genetic algorithm. The paper presents the results of experimental studies, confirming the effectiveness of the proposed hybrid intellectual technology for data classification.

Keywords: classifier based on the SVM algorithm, support vector, kernel function type, kernel function parameter, regularization parameter, classifier based on the Parzen windows, distance metric, hybrid technology, genetic algorithm

Введение. В настоящее время методы и алгоритмы интеллектуального анализа данных применяются для разработки разнообразных классификаторов данных. В качестве примера можно назвать классификаторы на основе деревьев решений; на основе искусственных нейронных сетей; на основе алгоритма опорных векторов – SVM-алгоритма (Support Vector Machine Algorithm); на основе алгоритма *k* ближайших соседей – *k*NN-алгоритма (*k* Nearest Neighbors Algorithm) и т.п. [1, 3, 7, 21]. Однако, как показывают результаты экспериментальных исследований, ни один из известных методов и алгоритмов, применяемых для разработки классификаторов данных, не может быть признан однозначно лучшим. Причина – ни один из них не обеспечивает гарантированно высокое качество классификации произвольных наборов данных [8, 16, 21, 23]. Этот факт объясняется спецификой математического инструментария, ограничивающей возможности разрабатываемого классификатора.

В последние годы для решения задач классификации активно используются классификаторы на основе SVM-алгоритма (KoSVM) [1, 3, 8–13, 20, 23]. Они обеспечивают высокое качество классификации сложноорганизованных многомерных данных. Однако такие классификаторы нередко дают ошибочные результаты вблизи гиперплоскости, разделяющей классы. В связи с этим очевидной является целесообразность поиска вспомогательного метода или алгоритма, который позволил бы адекватно решать задачу классификации данных именно вблизи гиперплоскости, разделяющей классы. Этой цели и посвящена данная работа.

Общая характеристика проблематики статьи. В ряде работ для гибридизации KoSVM предлагается использовать *k*NN-алгоритм. Подход к гибридизации, предложенный в [25], предполагает использование опорных векторов уже разработанного SVM-классификатора в качестве репрезентативного набора данных при разработке *k*NN-классификатора на основе *k*NN-алгоритма для классификации данных вблизи гиперплоскости, разделяющей классы. При этом утверждается, что при выполнении такой дополнительной классификации данных используется наиболее полезная информация. Однако использование этого подхода сопряжено с необходимостью адекватного определения числа соседей вновь классифицируемого объекта и может быть затруднено ввиду особенностей конфигурации опорных векторов в пространстве характеристик.

Основная идея гибридного подхода, предложенная в [16], заключается в применении локального KoSVM для классификации объекта, ошибочно классифицированного *k*NN-классификатором. При этом предлагается при разработке KoSVM использовать данные о ближайших соседях вновь классифицируемого объекта. Очевидно, что многократная разработка локальных KoSVM не является лучшим решением, так как сопряжена с больши*и*ми вычислительными затратами

Новая SVM-*k*NN-технология классификации данных была предложена одним из авторов этой работы в [4]. Эта технология демонстрирует повышение точности SVM-классификации для различных наборов данных посредством применения *k*NN-классификатора к данным в подобластях вблизи гиперплоскости, разделяющей классы.

По результатам анализа экспериментальных данных, полученных с применением SVM-*k*NN-технологии [4] с целью повышения точности классификации данных, выполняемой с применением KoSVM, было принято решение о поиске какого-либо нового вспомогательного метода или алгоритма для классификации данных в подобластях вблизи гиперплоскости, разделяющей классы. В качестве такого вспомогательного метода был выбран метод окна Парзена переменной ширины, хорошо зарекомендовавший себя при решении различных прикладных задач [1, 15, 18, 19]. Метод окна Парзена не так часто (возможно, ввиду некоторых особенностей установки значений своих параметров), как *k*NN-алгоритм, применяется в решении задач классификации. Однако при этом он демонстрирует высокое качество классификации данных и обладает некоторыми преимуществами над *k*NN-алгоритмом в отношении используемых правил при определении ближайших соседей для классифицируемого объекта [1]. При этом и *k*NN-классификатор и классификатор на основе метода окна Парзена (КоМОП) используют близкие по сути подходы к классификации данных.

В связи с этим было принято решение о разработке ГИТКД, основанной на совместном использовании SVM-алгоритма и метода окна Парзена. Предлагается применять КоМОП для классификации

объектов в определяемых экспериментально (по результатам использования KoSVM) подобластях пространства характеристик, расположенных вблизи границы классов и формирующих Ω-область, содержащую все объекты, ошибочно классифицированные с использованием KoSVM. При этом Ω-область может содержать некоторое число объектов, классифицированных правильно. Объекты, оказавшиеся за пределами Ω-области, предлагается использовать при разработке КоМОП как репрезентативные объекты соответствующих классов.

Пример классификации объектов на два класса с метками +1 и –1 в пространстве D-2 с применением KoSVM приведен на рисунке 1. При этом одна часть объектов, оказавшихся вне полосы, уверено разделяющей классы, классифицирована правильно. В то же время для другой части объектов, попавших внутрь разделяющей полосы и расположенных вблизи линии, определяющей границу классов, требуется применение дополнительного (вспомогательного) инструментария для определения принадлежности объектов к классам.

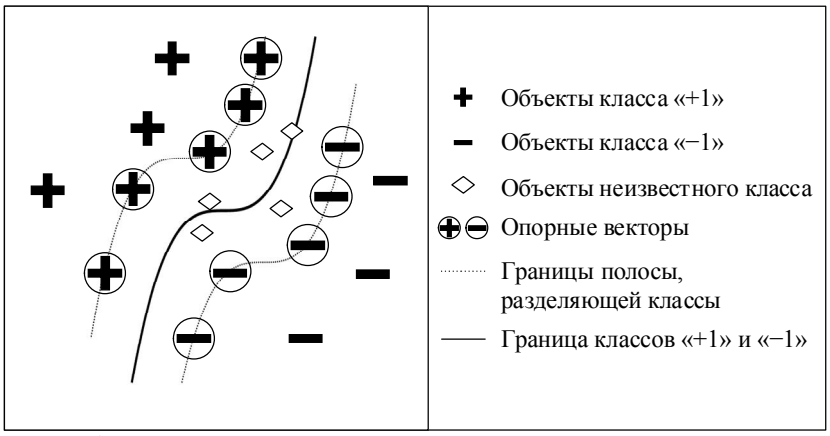


Рисунок 1 – Пример классификации объектов в пространстве D-2 с применением KoSVM

При реализации ГИТКД предлагается использовать KoSVM со значениями параметров, заданными по умолчанию, а субоптимальные значения параметров КоМОП определять с применением какого-либо эволюционного алгоритма оптимизации [2, 5, 6], например с применением генетического алгоритма. Использование такого подхода к определению значений параметров применяемых классификаторов позволит снизить временны́е затраты на получение гибридного классификатора, обеспечивающего высокую точность классификации данных.

Основные аспекты разработки классификатора на основе SVM-алгоритма. Пусть имеется набор данных вида: $\{(z_1,y_1),...,(z_s,y_s)\}$, в котором каждому объекту $z_i \in Z$ поставлено в соответствие число y_i, принадлежащее множеству меток классов $Y=\{-1;\ +1\}$ и задающее класс принадлежности объекта z_i, а сам объект z_1 представлен q-мерным вектором числовых значений характеристик:

$$z_i = (z_i^1, z_i^2, \ldots, z_i^q),$$

где z_i^l – значение l-ой характеристики для i-ого объекта ($i=\overline{1,s}$; $l=\overline{1,q}$) [14, 20, 23].

Рассматриваемый набор данных разбивается на две выборки: обучающую выборку Z^{train}, с использованием которой осуществляется обучение KoSVM; тестовую выборку Z^{test}, с использованием которой проводится дополнительная оценка качества классификационных решений, принимаемых обученным KoSVM. Если качество обучения и тестирования KoSVM является приемлемым, то он может быть рекомендован к применению для классификации новых объектов.

В случае бинарной классификации при обучении KoSVM на основе SVM-алгоритма для построения гиперплоскости, разделяющей классы, решается двойственная задача поиска седловой точки функции Лагранжа. Она сводится к задаче квадратичного программирования, содержащей только двойственные переменные [20]:

$$\begin{cases} -L(\lambda) = -\sum\limits_{i=1}^{S}\lambda_i + \\ \quad +\dfrac{1}{2}\cdot\sum\limits_{i=1}^{S}\sum\limits_{\tau=1}^{S}\lambda_i\cdot\lambda_\tau\cdot y_i\cdot y_\tau\cdot\kappa(z_i,z_\tau)\to\min\limits_{\lambda}, \\ \quad \sum\limits_{i=1}^{S}\lambda_i\cdot y_i = 0, \\ \quad 0\le\lambda_i\le C,\ i=\overline{1,S}, \end{cases} \tag{1}$$

где λ_i – двойственная переменная; z_i – объект из обучающей выборки; y_i – число (+1 или −1), характеризующее классовую принадлежность объекта z_i из обучающей выборки; $\kappa(z_i, z_\tau)$ – функция ядра; C – параметр регуляризации ($C > 0$); S – количество объектов в обучающей выборке ($S < s$); $i = \overline{1,S}$.

При этом в качестве функции ядра обычно использует линейную, радиальную базисную, полиномиальную или сигмоидную функции ядра.

В результате обучения KoSVM определяются опорные векторы – объекты из обучающей выборки, которые несут всю информацию о разделении классов и расположены ближе всего к гиперплоскости, разделяющей эти классы [3, 14, 20].

При разработке SVM-классификатора с помощью функции ядра $\kappa(z_i, z)$ формируется классифицирующая функция $f(z)$ [1, 14, 20]:

$$f(z) = \sum_{i=1}^{S} \lambda_i \cdot y_i \cdot \kappa(z_i, z) + b, \tag{2}$$

используемая для принятия решения о классовой принадлежности объекта z по правилу:

$$F(z) = sign(f(z)) = sign(\sum_{i=1}^{S} \lambda_i \cdot y_i \cdot \kappa(z_i, z) + b), \tag{3}$$

где b – смещение (bias).

Для обучения KoSVM необходимо определить значения его параметров: тип функции ядра $\kappa(z_i, z_\tau)$, значения параметров ядра и значение параметра регуляризации.

Обычно KoSVM, построенный даже со значениями своих параметров, заданными по умолчанию, обеспечивает высокое качество классификации данных. Для оценки качества классификации могут быть использованы различные показатели, такие как показатель общей точности (Overall Accuracy); показатель чувствительности (Sensitivity); показатель специфичности (Specificity); показатель точности (Precision); показатель сбалансированной F-меры (F-measure) [14, 23]. При этом большинство ошибочно классифицированных объектов попадают внутрь полосы, разделяющей классы и определяемой условием $-1 < < w, z > + b < 1$, где w − вектор-перпендикуляр к разделяющей гиперплоскости; z − вектор характеристик объекта; b − смещение, $< w, z >$ – скалярное произведение векторов w и z.

Очевидно, что для решения проблемы классификации объектов, находящихся внутри этой полосы, целесообразно использовать дополнительный инструментарий интеллектуального анализа данных.

Основные аспекты разработки классификатора на основе метода окна Парзена переменной ширины. Метод окна Парзена является частным случаем обобщенного метрического классификатора, заданного правилом классификации [1]:

$$a(u, Z^{train}) = \arg\max_{y \in Y} \sum_{i=1}^{\gamma} w_i [y_z^{(i)} = y], \tag{4}$$

где z – объект, классовую принадлежность которого необходимо установить; Z^{train} – обучающая выборка, состоящая из объектов, классовая принадлежность которых известна; w_i – вес объекта обучающей выборки; Y – множество меток классов; $y_z^{(i)}$ – классовая принадлежность i-ого соседа объекта z; γ – число объектов с меткой класса, равной y. Выражение $[y_z^{(i)} = y]$ принимает значение, равное «1», если выполняется условие в квадратных скобках, иначе – принимает значение «0».

Следует отметить, что обобщенный метрический классификатор позволяет работать с числом классов, равным двум или более.

В методе окна Парзена вес каждого объекта w_i задается с помощью некоторой функции ядра, невозрастающей в $[0, \infty)$ [19].

При реализации метода окна Парзена может использоваться окно фиксированной или переменной ширины. В соответствии с этим выбирается тот или иной способ задания веса объекта w_i.

Для окна Парзена фиксированной ширины вес объекта w_i вычисляется с учетом максимального расстояния h, на котором учитываются объекты:

$$w_i = K\left(\frac{d(z, z_z^{(i)})}{h}\right), \tag{5}$$

где K – функция ядра; h – положительное вещественное число; $d(z, z_z^{(i)})$ – расстояние от объекта z до своего i-ого соседа $z_z^{(i)}$, рассчитанное в соответствии с некоторой метрикой.

Для окна Парзена переменной ширины при определении веса объекта используется не ширина окна, а порядковый номер k для соседа $z_k^{(i)}$, вклад которого в классификацию объекта z учитывается:

$$w_i = K\left(\frac{d(z,z_z^{(i)})}{d(z,z_k^{(i)})}\right). \quad (6)$$

Таким образом, в формуле (6) вместо константы *h*, фигурирующей в знаменателе формулы (5), используется расстояние $d(z,z_k^{(i)})$ от объекта z до *k*-ого соседа из обучающей выборки Z^{train}.

Преимущество окна Парзена переменной длины заключается в возможности варьировать ширину окна так, чтобы окно всегда содержало ровно *k* ближайших соседей для объекта z. Поэтому при выполнении исследований будет использоваться именно окно Парзена переменной длины.

При разработке КоМОП на основе метода окна Парзена необходимо определить значения его параметров: число соседей *k*, метрику расстояния и функцию ядра.

Для вычисления расстояния между объектами $t=(t_1,t_2,\ldots,t_q)$ и $p=(p_1,p_2,\ldots,p_q)$ в *q*-мерном пространстве могут быть использованы такие метрики расстояний [17, 22]:

– Евклидово расстояние:

$$d(t,p)=\sqrt{\sum_{i=1}^{q}(t_i-p_i)^2}\ , \quad (7)$$

– квадрат Евклидова расстояния:

$$d(t,p)=\sum_{i=1}^{q}(t_i-p_i)^2\ , \quad (8)$$

– манхэттенское расстояние:

$$d(t,p)=\sum_{i=1}^{q}|t_i-p_i|\ , \quad (9)$$

– расстояние Чебышева:

$$d(t,p)=\max_i|t_i-p_i|\ . \quad (10)$$

В качестве функций ядра в методе окна Парзена могут быть использованы [1, 24]:

– функция ядра Епанечникова:

$$K(r)=\frac{3}{4}\cdot(1-r^2)\ ; \quad (11)$$

– квадратичная функция ядра:

$$K(r)=\frac{15}{16}\cdot(1-r^2)^2\ ; \quad (12)$$

– треугольная функция ядра:

$$K(r)=1-|r|\ ; \quad (13)$$

– гауссовская функция ядра:

$$K(r)=(2\pi)^{-\frac{1}{2}}\cdot e^{-\frac{1}{2}\cdot r^2}\ , \quad (14)$$

где *r* – вещественное число из отрезка [0; 1].

Предлагаемая технология классификации данных. Повышение качества классификации данных может быть достигнуто с использованием следующей ГИТКД.

1. Разработать KoSVM на основе исходного набора данных с использованием значений параметров классификатора, заданных по умолчанию. Оценить с применением различных показателей качество классификации данных.

2. Сформировать Ω-область, расположенную вблизи гиперплоскости, разделяющей классы, и содержащую все объекты, ошибочно классифицированные с применением KoSVM. При этом Ω-область может быть как симметричной, так и асимметричной. Асимметричная Ω-область может быть определена как $\Omega=\Omega^-\cup\Omega^+$, где Ω^- и Ω^+ – соответственно подобласти, содержащие все ошибочно классифицированные объекты, которые относятся к классам с метками –1 и +1 в исходном наборе данных. Симметричная Ω-область может быть определена как область, содержащая все ошибочно классифицированные объекты, расположенные на расстоянии, которое не превышает $\Delta=max\{d_{\Omega^-},\ d_{\Omega^+}\}$, где d_{Ω^-} – ширины подобласти Ω^-; d_{Ω^+} – ширины подобласти Ω^+. Сформированная на основе подобластей Ω^- и Ω^+ результирующая Ω-область будет включать в себя все ошибочно классифицированные объекты, образующие вместе с правильно классифицированными объектами, попавшими в Ω-область, и соответствующими метками классов объектов из Ω-области набор данных $G=\{<z_1,y_1>,\ldots,<z_{N_\Omega},y_{N_\Omega}>\}$. В этом наборе каждый кортеж $<z_i,y_i>$ содержит информацию об объекте z_i из Ω-области и соответствующую объекту z_i метку класса $y_i\in Y=\{-1;+1\}$.

3. Сформировать набор данных $V = U \setminus G$, который будет состоять только из тех кортежей исходного набора данных U, классовая принадлежность объектов для которых с помощью KoSVM была определена правильно.

4. Разработать КоМОП на основе набора данных $V = U \setminus G$ с применением ГА, используемого для поиска субоптимальных значений параметров КоМОП – числа ближайших соседей *k*, типа функции ядра и типа метрики расстояния – при классификации объектов из набора данных G.

5. Оценить качество итоговой классификации данных с применением различных показателей качества.

Совместное использование вышеуказанных классификаторов позволит гармонично сочетать их достоинства и нивелировать недостатки. Высокие вычислительные затраты на разработку KoSVM могут быть компенсированы существенно меньшими временны́ми затратами на разработку КоМОП. При этом при разработке КоМОП придется иметь дело уже не со всем исходным набором данных, а с набором существенно меньшей мощности, содержащим только объекты, которые расположены за пределами Ω-области.

Использование нового дополнительного инструментария – КоМОП, основанного на других принципах интеллектуального анализа данных (по сравнению с KoSVM) позволит повысить общую точность классификации данных. При этом предлагаемая ГИТКД может быть применена к классификации новых объектов, если она обеспечивает повышение качества классификации на исходном наборе данных.

Ограничение применимости предлагаемого метода классификации заключается в следующем: если ширина Ω-области очень велика или плотность объектов внутри Ω-области чрезмерна, то число объектов может оказаться недостаточным для последующей разработки КоМОП.

Использование генетического алгоритма для настройки параметров КоМОП. Вообще говоря, поиск оптимальных значений параметров КоМОП может быть реализован посредством перебора всех возможных сочетаний значений числа ближайших соседей *k*, типа функции ядра и типа метрики расстояния. Однако такой поиск может сопровождаться ощутимыми временны́ми затратами при увеличении числа типов функций ядра, числа метрик расстояний и особенно – числа возможных соседей.

Для сокращения временных затрат на поиск оптимальных значений параметров КоМОП предлагается использовать ГА, оперирующий целочисленными значениями параметров. В общем случае он обеспечивает получение не оптимального, а субоптимального решения. Пусть, например, число возможных ближайших соседей *k* выбрано так, что удовлетворяет условиям: $1 \le k \le N/3$; k – целое нечетное число (это позволяет избежать проблем с голосованием), где N – число объектов в обучающем наборе данных, сформированном на шаге 3 ГИТКД (см. предыдущий раздел). Пронумеруем типы метрик расстояния (7) – (10) целыми числами от 1 до 4, а типы функций ядра (11) – (14) целыми числами от 1 до 4. Тогда *m*-ая хромосома в ГА может быть закодирована как:

$$v = (v_1, v_2, v_3) \tag{15}$$

где v_1 – число ближайших соседей; v_2 – номер типа метрики расстояния; v_3 – номер типа функции ядра.

В качестве фитнесс-функции ГА может быть использован тот или иной показатель качества классификации.

В исследовании была использована стандартная схема реализации ГА:

1. Выполнить инициализацию популяции хромосом размером P.

2. Выполнить расчет «успешности» каждой хромосомы с использованием фитнесс-функции. Проверить условие останова алгоритма. Если условие останова выполняется, то завершить работу, иначе перейти к шагу 3.

3. Выполнить генетический оператор селекции с целью отбора наиболее успешных хромосом на роль родителей для образования на основе их генов поколения детей.

4. Создать поколение детей с применением генетических операторов скрещивания и мутации. В зависимости от типа применяемого оператора детские особи могут быть произведены от одного родителя случайным изменением его генов (в случае применения оператора мутации) или от двух родителей комбинированием их генов (в случае применения оператора скрещивания).

5. Сформировать результирующее поколение хромосом размером P, используя поколение родителей и поколение детей с соблюдением принципа элитизма.

В качестве критерия останова ГА можно использовать следующие варианты: критерий достижения максимального числа поколений; критерий достижения максимального времени выполнения алгоритма; критерий превышения числа застойных поколений. Застойным поколением можно считать поколение хромосом, лучшее значение целевой функции которого не отличается на некоторую малую константу ε ($\varepsilon > 0$) от лучшего значения фитнесс-функции предыдущего поколения.

Особый интерес при реализации ГА вызывают варианты задания генетических операторов: селекции, скрещивания и мутации, а также, например, варианты задания значений параметров операторов скрещивания и мутации (вероятностей этих событий).

В рамках решаемой задачи не было необходимости в использовании избыточно сложных и самонастраивающихся операторах скрещивания и мутации ввиду дискретности значений параметров фитнесс-

функции. В связи с этим были использованы стандартные варианты задания генетических операторов: в случае оператора мутации к значениям случайных генов хромосомы родителя добавляются случайные числа из распределения Гаусса; в случае оператора скрещивания – детская хромосома случайным образом наследует ген одного из родителей.

Рассматриваемый ГА может быть использован после соответствующей адаптации для одновременного поиска лучшего варианта вспомогательного классификатора из некоторого набора потенциально возможных классификаторов и значений параметров лучшего вспомогательного классификатора.

В качестве потенциально возможных вспомогательных классификаторов могут использоваться, например, *k*NN-классификатор [4] и КоМОП. Эффективность их применения при реализации ГИТКД уже доказана. В дальнейшем набор потенциально возможных вспомогательных классификаторов может быть расширен.

На рисунке 2 представлена укрупненная схема работы предлагаемой ГИТКД на примере смоделированного набора данных U, содержащего 15 объектов класса +1 (обозначены на рисунке как знак плюс) и 13 объектов класса –1 (обозначены на рисунке как знак минус). Объекты набора данных U были разделены на обучающую и тестовую выборки, содержащие соответственно объекты с заливкой и без заливки (рис. 2).

Для набора данных U согласно шагу 1 ГИТКД производится разработка KoSVM со значениями параметров, заданными по умолчанию. На этом шаге для рассматриваемого набора данных U определено 10 опорных векторов, которые задают информацию о разделении классов. При этом три объекта из восьми классифицированы неверно (это объекты, заключенные в рамку на рис. 2).

Согласно шагу 2 определяются симметричные и асимметричные Ω -области, содержащие ошибочно классифицированные объекты.

На шаге 3 формируются наборы данных V, из которых исключена информация обо всех объектах, расположенных внутри симметричной и асимметричной областей. Наборы данных V используются в качестве новых наборов данных для разработки КоМОП. В рассматриваемом примере из исходного набора данных U исключена информация о семи и шести объектах соответственно для вариантов с симметричной и асимметричной Ω -областями.

Построенный на шаге 4 КоМОП с использованием ГА позволяет верно классифицировать объекты как симметричной, так и асимметричной Ω -областей.

На шаге 5 выбирается лучший из разработанных КоМОП, обеспечивающий минимизацию числа ошибок при классификации данных. Затем составляется новое правило классификации, условно обозначенное на рисунке 2 как преобразованная с учетом КоМОП разделяющая гиперплоскость, сформированная при разработке KoSVM.

Результаты вычислительных экспериментов. Апробация ГИТКД производилась на реальных наборах данных, взятых из проекта Statlog и репозитория задач машинного обучения UCI Machine Learning Repository. В частности, были использованы наборы данных медицинской диагностики (WDBC и Heart; источники http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/ и http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/ соответственно), кредитного скоринга (Firms и German, источники [10] и http://archive.ics.uci.edu/ml/machine-learning-databases/ statlog/german/), обработки сигналов (Ionosphere, источник https://archive.ics.uci.edu/ml/machine-learning-databases/ionosphere/) и набор модельных данных (MOTP12, источник http://machinelearning.ru/wiki/images/b/b2/MOTP12_ svm_example.rar) (табл. 1). На всех наборах данных использовался вариант бинарной классификации.

Экспериментальные исследования (вычислительные эксперименты) производились с использованием ПЭВМ, работающей под 64-разрядной версией “Windows 10” с оперативной памятью 8 Гб и двухъядерным процессором “Intel® Core™ i3-4160” с тактовой частотой каждого ядра 3,6 ГГц. В ходе исследований использовалась программная реализация SVM-алгоритма, предоставляемая системой инженерных и научных расчетов “MATLAB 7.12.0.635”.

Для каждого набора данных было произведено его многократное разбиение на обучающую и тестовую выборки с последующей разработкой KoSVM. Разработка KoSVM выполнялась со значениями параметров, заданными по умолчанию: значением параметра регуляризации $C = 1$; радиальной базисной функции ядра со значением параметра $\sigma = 1$. В качестве лучшего KoSVM выбирался тот, который обеспечивал бóльшее значение общей точности классификации на обучающей и тестовой выборках. При этом выполнялся анализ числа ошибок на обучающей и тестовой выборках.

По результатам разработки KoSVM осуществлялся анализ возможности применения КоМОП. В частности, выполнялась оценка числа объектов, попавших внутрь разделительной полосы: $-1 < < w, z > + b < 1$. Для всех наборов данных, использовавшихся в эксперименте, оказалось, что все ошибочно классифицированные объекты расположены внутри разделительной полосы.

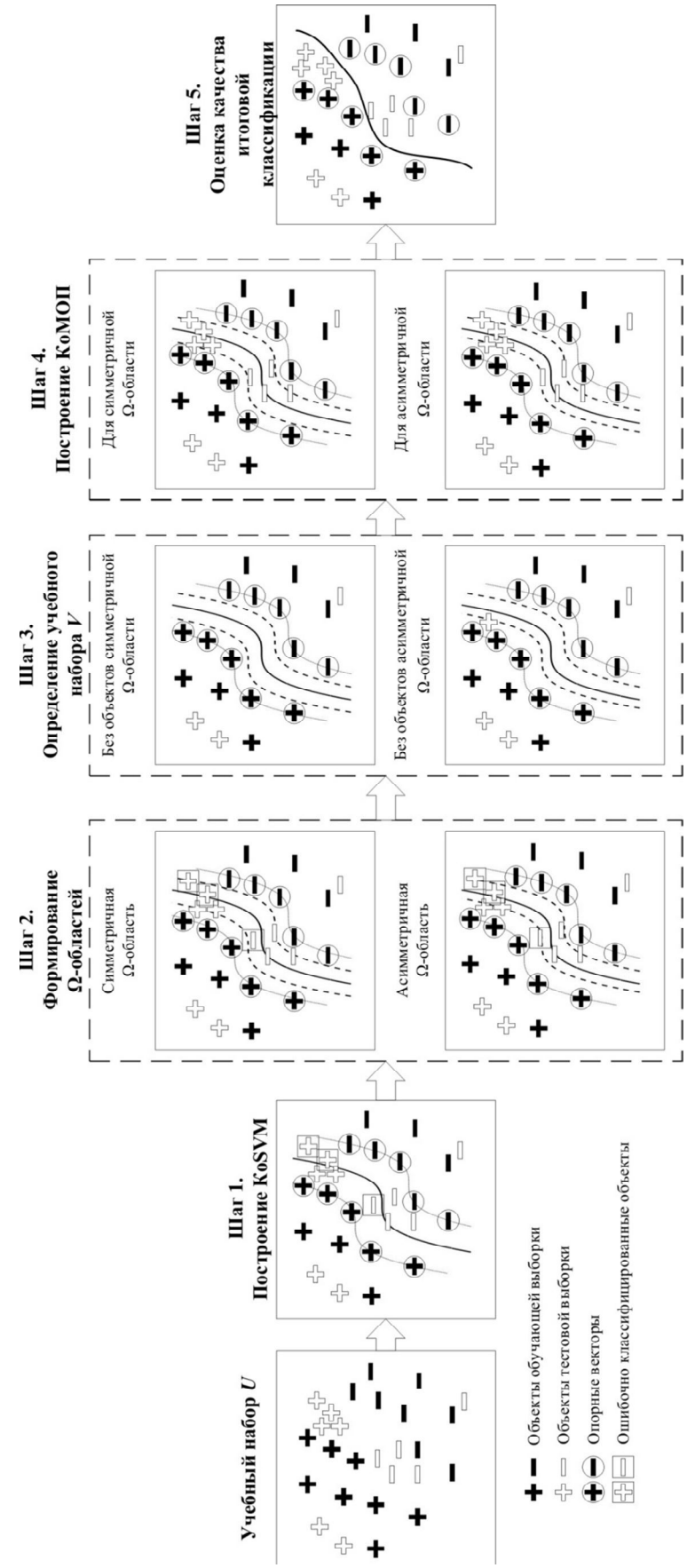


Рисунок 2 – Укрупненная схема работы гибридной интеллектуальной технологии классификации данных

Кроме того, выполнялась оценка ширина Ω-области, в которую попадают все ошибочно классифицированные с применением KoSVM объекты, и число объектов внутри Ω-области. Для всех наборов данных, использовавшихся в эксперименте, оказалось, что Ω-область расположена внутри разделительной полосы.

Для всех исходных наборов данных число объектов, оказавшихся в новых наборах данных, полученных после удаления из исходных информации об объектах из Ω-области, оказалось достаточным для разработки соответствующих КоМОП. Поэтому был выполнен поиск субоптимальных значений параметров этих КоМОП (числа соседей k, метрики расстояния и функции ядра) с применением ГА.

В рамках экспериментов размер популяции ГА составлял 20 хромосом; число застойных поколений было выбрано равным 20; минимальное отклонение значения фитнесс-функции для определения застойного поколения – $\varepsilon = 1 \cdot 10^{-10}$; максимальное число поколений – 200; вероятность скрещивания – 0,8; вероятность мутации – 0,2.

Для обоснования предпочтительности ГА полному перебору параметров при разработке КоМОП, можно привести следующий пример. Для набора данных МОТП12 (http://machinelearning.ru/wiki/images/b/b2/MOTP12_ svm_example.rar), состоящего из 400 объектов, с двумя характеристиками, полный перебор значений параметров КоМОП требует выполнения разработки 864 КоМОП. Это составляет 12 с в случае, когда используются 4 функции ядра, 4 метрики расстояний, а число вариантов для числа k ближайших соседей равно 54 (число ближайших соседей удовлетворяет условию $1 \le k \le [320/3] = 107$, где k – нечетное число). В ходе вычислительного эксперимента исходный набор данных МОТП12 случайным образом разбивался на обучающую и тестовую выборки, содержавшие соответственно 320 и 80 объектов.

Разработка КоМОП подразумевает следующее: расчет расстояния от объекта неизвестной классовой принадлежности до каждого объекта обучающей выборки; сортировку объектов обучающей выборки в соответствии с рассчитанным расстоянием; определение классовой принадлежности неизвестного объекта с учетом весов соседей.

В случае использования ГА при поиске субоптимальных значений параметров КоМОП при 50 запусках ГА среднее время поиска составляет 3 с для популяции из 20 хромосом и числе застойных поколений, равном 5. При этом среднее число разработанных КоМОП равно 185. Таким образом, время поиска субоптимальных значений параметров КоМОП уменьшилось в 4 раза по сравнению с вариантом полного перебора вариантов.

При этом следует отметить, что ГА является эвристическим, и он в общем случае не гарантирует сходимость за определенное число шагов. При работе со сложноорганизованными наборами данных большой размерности практически невозможно предсказать поведение фитнесс-функции ГА, в качестве которой может выступать тот или иной показатель качества классификации, например показатель общей точности, F-мера и т.п. [23]. Однако при решении многих оптимизационных задач, когда неизвестна специфика поведения фитнесс-функции, ГА позволяет получить приемлемые (удовлетворительные) результаты, которые принимаются в качестве субоптимальных. Для предотвращения слишком долгого решения оптимизационной задачи с применением ГА устанавливается ограничение на максимальное число поколений.

Специфика решаемой задачи заключается еще и в том, что при реализации различных прогонов ГИТКД для одного и того же исходного набора данных возможно получение различных обучающих и тестовых выборок. Соответственно, будут иметь место различные вариантов наборов данных V, содержащие объекты из Ω-области, и различные варианты решения задачи поиска значений параметров КоМОП.

Целесообразность использования ГА может быть также обоснована планируемой в дальнейшем его адаптацией к реализации одновременного поиска лучшего варианта вспомогательного классификатора из некоторого набора потенциально возможных классификаторов (типа КоМОП, kNN-классификатора и их модификаций) и значений параметров лучшего вспомогательного классификатора.

В столбце 1 таблицы 1 содержится название исходного набора данных, число объектов и число характеристик у каждого объекта классификации. Столбцы 2–4 объединяют информацию о результатах использования KoSVM: число опорных векторов и число ошибок на каждой выборке. В столбце 5 указан вариант использования классификаторов (соответственно только KoSVM – SVM, KoSVM и КоМОП с ассиметричной областью – aSim, KoSVM и КоМОП с симметричной областью – Sim). Столбцы 6 и 7 указывают ширину Ω-области и число объектов, лежащих в этой области. Найденные с применением ГА значения параметров КоМОП отображены в столбцах 8–10. Число ошибок на обучающем и тестовом наборах указаны в столбцах 11–12. В 13 столбце приведены значения показателя общей точности (Overall accuracy) [23].

Для набора данных WDBC (http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/), содержащего 569 объектов, разработка KoSVM со значениями параметров, заданными по умолчанию, привела к 0 ошибок на обучающей выборке и 42 ошибкам на тестовой выборке. При этом ошибочно классифицированные объекты находятся с одной стороны от разделяющей полосы на расстоянии 0.373 (в то время как опорные векторы находятся на расстоянии, равном 1, от разделяющей гиперплоскости). Симметричная и асимметричная Ω-области содержат равное число объектов – 42, поэтому

их использование дает одинаковые результаты при разработке КоМОП. Поиск субоптимальных значений параметров КоМОП с применением ГА дал такие результаты: число ближайших соседей, равное 3; целесообразно использование Евклидовой метрики расстояния и функции ядра Епанечникова. Применение предлагаемой ГИТКД позволило избавиться от ошибок на тестовой выборке.

Таблица 1 – Значения параметров классификаторов и значения показателей качества классификации

Набор данных	SVM классификация			Вариант классификации	Ω		Классификация на основе метода окна Парзена			Число ошибок		Общая точность, %
	Число опорных векторов	Число ошибок										
		Er_{train}	Er_{test}		d_Ω	N_Ω	h	Метрика расстояния	Функция ядра	Er_{train}	Er_{test}	
1	2	3	4	5	6	7	8	9	10	11	12	13
WDBC (569 × 30)	455	0	42	SVM	—	—	—	—	—	0	42	92,62
		из	из	+aSim	0,373	42	3	Евкл. р.	Епанешн.	0	0	100,00
		456	113	+Sim	0,745	42	3	Евкл. р.	Епанешн.	0	0	100,00
German (1000 × 24)	796	0	60	SVM	—	—	—	—	—	0	60	94,00
		из	из	+aSim	0,784	199	137	р. Чеб.	Квадратичн.	1	48	95,10
		800	200	+Sim	1,567	434	-	—	—	—	—	—
Heart (270 × 13)	216	0	20	SVM	—	—	—	—	—	0	20	92,59
		из	из	+aSim	0,138	37	23	р. Чеб.	Епанешн.	2	2	98,52
		216	54	+Sim	0,276	38	33	р. Чеб.	Гауссовская	2	2	98,52
MOTP12 (400 × 2)	132	32	1	SVM	—	—	—	—	—	32	1	91,75
		из	из	+aSim	2,000	97	5	р. Чеб.	Епанешн.	26	2	93,00
		360	40	+Sim	2,000	97	5	р. Чеб.	Епанешн.	26	2	93,00
Firms (60 × 12)	46	0	6	SVM	—	—	—	—	—	0	6	90,00
		из	из	+aSim	0,027	10	11	Манхэт. р.	Квадратичн.	0	0	100,00
		48	12	+Sim	0,055	10	11	Манхэт. р.	Квадратичн.	0	0	100,00
Ionosphere (351 × 34)	249	0	22	SVM	—	—	—	—	—	0	22	93,73
		из	из	+aSim	0,020	23	25	Манхэт. р.	Квадратичн.	0	1	99,72
		281	70	+Sim	0,041	23	25	Манхэт. р.	Квадратичн.	0	1	99,72

Для набора данных German (http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/), содержащего 1000 объектов, число ошибок на обучающей и тестовой выборках при разработке KoSVM со значениями параметров, заданными по умолчанию, привела, соответственно, к 0 и 60 ошибкам. Использование симметричной Ω-области оказалось нецелесообразным, так как число исключаемых объектов составляет бóльшую часть от количества объектов в исходном наборе данных. Поэтому вследствие исключения из исходного набора данных информации об объектах из Ω-области классы становятся существенно несбалансированными.

Исключение из исходного набора данных объектов из ассиметричной Ω-области шириной 0,784, содержащей 199 объектов, позволило получить при разработке КОМОП число ошибок на обучающей и тестовой выборках, равное 1 и 48 соответственно. Увеличение числа ошибок на обучающей выборке можно объяснить возможным исключением опорных векторов, попавших в ассиметричную Ω-область, из исходного набора данных. Поиск субоптимальных значений параметров КоМОП с применением ГА дал такие результаты: число ближайших соседей равно 137; целесообразно использование метрики расстояние Чебышева и функции ядра Епанечникова.

Для набора данных Heart (http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/), содержащего 270 объектов, разработка KoSVM со значениями параметров, заданными по умолчанию, привела к 0 и 20 ошибкам на обучающей и тестовой выборках соответственно. При этом оказалось, что асимметричная и симметричная Ω-области содержат соответственно 37 и 38 объектов. Поиск субоптимальных значений параметров классификатора на основе окна Парзена с применением ГА дал следующие результаты: число ближайших соседей равно 33; целесообразно использование метрики Чебышева и гауссовской функции ядра. Число ошибок при применении предлагаемой ГИТКД составило 2 и 2 на обучающей и тестовой выборках соответственно.

Для набора данных MOTP12 (http://machinelearning.ru/wiki/ images/b/b2/MOTP12_ svm_example.rar), содержащего 400 объектов, разработка KoSVM со значениями параметров, заданными по умолчанию, привела к 32 и 1 ошибкам на обучающей и тестовой выборках соответственно. Так как для этого набора характерно сложное разделение на классы, то по результатам разработки KoSVM со

значениями параметров, заданными по умолчанию, оказалось, что большое число ошибочно классифицированных объектов лежит вне разделяющей полосы, причем их расстояния от этой полосы больше 1. Во избежание избыточного исключения информации из исходного набора данных было принято решение об исключении информации только о тех объектах, которые попали внутрь разделяющей полосы. В результате ширина как симметричной, так и ассиметричной Ω-области оказалась равна 2. При этом в Ω-область попало 97 объектов. Поиск субоптимальных значений параметров КоМОП с применением ГА дал такие результаты: число ближайших соседей равно 5; целесообразно использование метрики расстояние Чебышева и функции ядра Епанечникова. Число ошибок при применении предлагаемой технологии классификации данных составило 26 и 2 на обучающей и тестовой выборках соответственно.

Для набора данных Firms [10], содержащего 60 объектов, разработка KoSVM со значениями параметров, заданными по умолчанию, привела к 0 и 6 ошибкам на обучающей и тестовой выборках соответственно. При этом оказалось, что симметричная и асимметричная Ω-области содержат одинаковое число объектов, равное 10, следовательно, их использование дает одинаковые результаты при разработке КоМОП. Поиск субоптимальных значений параметров КоМОП с применением ГА дал следующие результаты: число ближайших соседей равно 11; целесообразно использование манхэттеновской метрики и квадратичной функции ядра. Использование предлагаемой ГИТКД позволило уменьшить число ошибок на обучающей и тестовой выборках до 0.

Для набора данных Ionosphere (https://archive.ics.uci.edu/ml/machine-learning-databases/ionosphere/), содержащего 351 объект, разработка KoSVM со значениями параметров по умолчанию, привела к 0 и 22 ошибкам на обучающей и тестовой выборках соответственно. Для этого набора данных, как и для набора WDBC, оказалось, что все ошибочно классифицированные объекты лежат по одну сторону от разделяющей гиперплоскости. При этом симметричная и асимметричная Ω-области содержат одинаковое число объектов, равное 23 (так как всего ошибочно классифицированных объектов – 22, то в Ω-область попал один объект, правильно классифицированный с применением KoSVM). Поиск субоптимальных значений параметров КоМОП с применением ГА дал такие результаты: число ближайших соседей равное 24; целесообразность использования метрики манхэттеновского расстояния и квадратичной функции ядра. Число ошибок при применении предлагаемой ГИТКД на обучающей и тестовой выборках составило 0 и 1 соответственно.

Был проведен сравнительный анализ результатов классификации рассмотренных в настоящей работе наборов данных с применением KoSVM в качестве основного классификатора и КоМОП в качестве вспомогательного классификатора и результатов классификации, полученных в [4] при применении *k*NN-классификатора в качестве вспомогательного. Результаты сравнительного анализа (табл. 2) позволяют сделать ряд выводов.

Использование КоМОП (вместо *k*NN-классификатора) в качестве вспомогательного классификатора в дополнение к KoSVM позволило обеспечить следующее: получить аналогичные значения ошибок на обучающей и тестовой выборках для наборов Firms и WDBC; улучшить результаты классификации наборов данных Heart и Ionosphere. При этом для набора данных German использование КоМОП (вместо *k*NN-классификатора) в качестве вспомогательного классификатора в дополнение к KoSVM привело к ухудшению результатов классификации. Приведенные результаты вычислительных экспериментов демонстрируют целесообразность применения предлагаемой ГИТКД, предполагающей последовательное использование KoSVM и КоМОП.

В качестве возможных вариантов улучшения результатов классификации с применением рассматриваемой ГИТКД целесообразно указать следующие подходы: использование при разработке КоМОП взвешенных метрик как способа отбора признаков и реализации механизма обучения; поиск ближайших соседей с учетом взаимного расположения объектов обучающей выборки, а не посредством пересчета расстояний до всех объектов; отбор наиболее информативных объектов обучающей выборки с помощью алгоритмов отбора эталонов для метрических классификаторов.

Таблица 2. Сравнение результатов классификации с применением различных вспомогательных классификаторов, используемых в дополнение к KoSVM

Набор	SVM+*k*NN			SVM+Parzen			Примечание
	Число ошибок			Число ошибок			
	Er_{train}	Er_{test}	*Sum*	Er_{train}	Er_{test}	*Sum*	
Firms	0	0	0	0	0	0	Без изменений
WDBC	0	0	0	0	0	0	Без изменений
German	12	29	41	1	48	49	Ухудшение
Heart	2	3	5	2	2	4	Улучшение
Ionosphere	1	1	2	0	1	1	Улучшение

Заключение. По результатам проведенных вычислительных экспериментов можно сделать вывод о том, что использование предлагаемой ГИТКД повышает качество результатов классификации, поскольку применение КоМОП к объектам, расположенным вблизи разделяющей гиперплоскости, определенной с помощью KoSVM, уменьшает число ошибочно классифицированных объектов.

Использование КоМОП в качестве вспомогательного классификатора вместо *k*NN-классификатора можно обосновать тем, что *k*NN-классификатор использует расстояние до объекта для определения ближайших соседей, но при этом не использует это расстояние при принятии решения о классовой принадлежности объекта. Определение функции весов в зависимости от порядкового номера ближайшего соседа в некоторых модификациях *k*NN нецелесообразна. Причина – такой подход не позволяет оценить фактическое расположение объектов в пространстве характеристик. КоМОП принимает решение о классовой принадлежности объекта на основе расстояния до соседа с помощью функции ядра.

Целью дальнейших исследований будет являться адаптация ГА к реализации одновременного поиска лучшего варианта вспомогательного классификатора (ЛВВК) из некоторого набора потенциально возможных классификаторов и значений параметров для ЛВВК.

Библиографический список

1. Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин) / К. В. Воронцов. – 141 с. – Режим доступа: www.MachineLearning.ru, свободный. – Заглавие с экрана. – Яз. рус. (Дата обращения: 01.03.2018).

2. Гуменникова А. В. Адаптивные поисковые алгоритмы для решения сложных задач многокритериальной оптимизации : дис. … канд. тех. наук / А. В. Гуменникова. – Красноярск, 2006. – 129 с.

3. Демидова Л. А. Использование модифицированного алгоритма роя частиц в задаче разработки SVM-классификатора / Л. А. Демидова, Ю. С. Соколова //Прикаспийский журнал: управление и высокие технологии. – 2016. – № 1 (33). – С. 26–38 (http://hi-tech.asu.edu.ru/files/1(33)/26-38.pdf).

4. Демидова Л. А. Классификация данных на основе SVM алгоритма и алгоритма k-ближайших соседей данных / Л. А. Демидова, Ю. С. Соколова // Вестник Рязанского государственного радиотехнического университета. – 2017. – № 62. – С. 119–132.

5. Семенкин Е. С. Адаптивные поисковые методы оптимизации сложных систем / Е. С. Семенкин, О. Э. Семенкина, С. П. Коробейников. – Красноярск : СИБУП, 1997. – 355 с.

6. Сопов Е. А. Эволюционные алгоритмы моделирования и оптимизации сложных систем : дис. … канд. тех. наук / Е. А. Сопов. – Красноярск, 2004. – 118 с.

7. Хлопкова О. Нейроэволюционный метод интеллектуализации принятия решений в условиях неопределенности / О. Хлопкова // Прикаспийский журнал: управление и высокие технологии. – 2015. – № 3. – С. 114–129 (http://hi-tech.asu.edu.ru/files/3(31)/114-129.pdf).

8. Хритоненко Д. И. Адаптивные коллективные нейро-эволюционные алгоритмы интеллектуального анализа данных : дис. … канд. тех. наук / Д. И. Хритоненко. – Красноярск, 2017. – 126 с.

9. Akhmedova Sh. A. SVM-based classifier ensembles design with co-operative biology inspired algorithm / Sh. A. Akhmedova, Ye. S. Semenkin // Вестник СибГАУ. – 2015. – № 1 (16). – P. 22–27.

10. Beynon M. J. Variable precision rough set theory and data discretisation: an application to corporate failure prediction / M. J. Beynon, M. J. Peel // Omega. – 2001. – Vol. 29. – P. 561–576.

11. Chapelle O. Choosing Multiple Parameters for Support Vector Machine / O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee // Machine Learning. – 2002. – № 46 (1–3). – P. 131–159.

12. Demidova L. Use of Fuzzy Clustering Algorithms' Ensemble for SVM classifier Development / L. Demidova, Yu. Sokolova, E. Nikulchev // International Review on Modelling and Simulations (IREMOS). – 2015. – Vol. 8, № 4. – P. 446–457.

13. Demidova L. Modification of Particle Swarm Algorithm for the Problem of the SVM Classifier Development / L. Demidova, Yu. Sokolova // International Conference «Stability and Control Processes» in Memory of V. I. Zubov. – 2015. – P. 623–627.

14. Demidova L. A. Big Data Classification Using the SVM Classifiers with the Modified Particle Swarm Optimization and the SVM Ensembles / L. A. Demidova, E. V. Nikulchev, Yu. S. Sokolova // International Journal of Advanced Computer Science and Applications (IJACSA). – 2016. – Vol. 7, no. 5. – P. 294–312.

15. Jain A. K. Classifier Design with Parzen Windows / A. K. Jain, M. D. Ramaswami // Pattern Recognition and Artificial Intelligence. Elsevier Sci. Publishers. – 1988. – Vol. 7. – P. 211–228.

16. Li R. Support Vector Machine combined with K-Nearest Neighbors for Solar Flare Forecasting / R. Li, H.-N. Wang, H. He, Y.-M. Cui, Zh.-L. Du // Chinese Journal of Astronomy and Astrophysics. – 2007. – Vol. 7, № 3. – P. 441–447.

17. Mousa A. An improved Chebyshev distance metric for clustering medical images / A. Mousa, Yu. Yusof // AIP Conference Proceedings. – 2015. – Vol. 1691, issue 1. – P. 040020.

18. Pan Zh.-W. Parzen windows for multi-class classification / Zh.-W. Pan, D.-H. Xiang, Q.-W. Xiao, D.-X. Zhou // Journal of Complexity. – 2008, October – December. – Vol. 24, issues 5–6. – P. 606–618.

19. Parzen E. On Estimation of 3 Probability Density Function and Mode / E. Parzen // The Annals of Mathematical Statistics. – 1962. Vol. 33, № 3. – P. 1065–1076.

20. Vapnik V. Statistical Learning Theory / V. Vapnik. – New York : John Wiley & Sons, inc., 1998. – P. 740.

21. Wang H. Extended k-Nearest Neighbours Based on Evidence Theory / H. Wang, D. Bell // The Computer Journal. – 2004. – Vol. 47 (6). – P. 662–672.

22. Weinberger K. Q. Distance Metric Learning for Large Margin Nearest Neighbor Classification / K. Q. Weinberger, L. K. Saul // Journal of Machine Learning Research. – 2009. – Vol. 10. – P. 207–244.

23. Yu L. Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines / L. Yu, Sh. Wang, K. K. Lai, L. Zhou. – Springer-Verlag Berlin Heidelberg, 2008. – P. 244.

24. Zambom A. Z. A Review of Kernel Density Estimation with Applications to Econometrics / A. Z. Zambom, R. Dias // International Econometric Review (IER), Econometric Research Association. – 2013. – Vol. 5 (1). – P. 20–42.

25. Zhang H. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition / H. Zhang, A. C. Berg, M. Maire. J. Malik // Proceedings – 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. – 2006. – Vol. 2. – P. 2126–2136.

References

1. Vorontsov K. V. *Matematicheskie metody obucheniya po pretsedentam (teoriya obucheniya mashin)* [Mathematical methods of training in precedents (theory of machine training). 141 p., Available at: www.MachineLearning.ru (Accessed: 01.03.2018).

2. Gumennikova A. V. *Adaptivnye poiskovye algoritmy dlya resheniya slozhnykh zadach mnogokriterialnoy optimizatsii: Dissertatsiya* [Adaptive search algorithms to solve complex problems multi-objective optimization: Dissertation]. Krasnoyarsk, 2006, 129 p.

3. Demidova L. A., Sokolova Yu. S. Ispolzovanie modifitsirovannogo algoritma roya chastits v zadache razrabotki SVM-klassifikatora [The use of the modified particle swarm algorithm in the development problem of the SVM classifier]. *Prikaspijskiy jurnal: upravlenie i vysokie tehnologii* [Caspian Journal: Management and High Technologies], 2016, no. 1 (33), pp. 26–38 (http://hi-tech.asu.edu.ru/files/1(33)/26-38.pdf).

4. Demidova L. A., Sokolova Yu. S. Klassifikatsiya dannykh na osnove SVM algoritma i algoritma k-blizhayshikh sosedey dannykh [Data Classification based on the SVM Algorithm and the k-Nearest-Neighbor Algorithm]. *Vestnik Ryazanskogo gosudarstvennogo radiotekhnicheskogo universiteta* [Bulletin of Ryazan State Radio Engineering University], 2017, no. 62, pp. 119–132.

5. Semenkin Ye. S., Semenkina O. E., Korobeynikov S. P. *Adaptivnye poiskovye metody optimizatsii slozhnykh sistem* [Adaptive search methods for optimization of complex systems]. Krasnoyarsk, SIBUP Publ., 1997, pp. 355.

6. Sopov Ye. A. *Evolyutsionnye algoritmy modelirovaniya i optimizatsii slozhnykh system: Dissertatsiya* [Evolutionary algorithms for simulation and optimization of complex: Dissertation]. Krasnoyarsk, 2004, 118 p.

7. Khlopkova O. Neyroyevolyutsionnyy metod intellektualizatsii prinyatiya resheniy v usloviyakh neopredelennosti [Neuro-evolutionary method of intellectualization of decision-making under conditions of uncertainty]. *Prikaspiyskiy: upravlenie i vysokie tehnologii* [Caspian Journal: Management and High Technologies], 2015, no. 3, pp. 114–129 (http://hi-tech.asu.edu.ru/files/3(31)/114-129.pdf).

8. Khritonenko D. I. *Adaptivnye kollektivnye neyro-ehvolyutsionnye algoritmy intellektualnogo analiza dannykh: Dissertatsiya* [Adaptive collective neuro-evolutionary algorithms for data mining: Dissertation]. Krasnoyarsk, 2017, 126 p.

9. Akhmedova Sh. A., Semenkin E. S. SVM-based classifier ensembles design with co-operative biology inspired algorithm. *Bulletin of Siberian State University of Management*, 2015. no. 1 (16), pp. 22–27.

10. Beynon M. J., Peel M. J. Variable precision rough set theory and data discretisation: an application to corporate failure prediction. *Omega*, 2001, vol. 29, pp. 561–576.

11. Chapelle O., Vapnik V., Bousquet O., Mukherjee S. Choosing Multiple Parameters for Support Vector Machine. *Machine Learning*, 2002, no. 46 (1–3), pp. 131–159.

12. Demidova L., Nikulchev E., Sokolova Yu. Use of Fuzzy Clustering Algorithms' Ensemble for SVM classifier Development. *International Review on Modelling and Simulations (IREMOS)*, 2015, vol. 8, no. 4, pp. 446–457.

13. Demidova L, Sokolova. Yu. Modification of Particle Swarm Algorithm for the Problem of the SVM Classifier Development. *International Conference "Stability and Control Processes" in Memory of V.I. Zubov*, 2015, pp. 619–622.

14. Demidova L. A., Nikulchev E. V., Sokolova Yu. S. Big Data Classification Using the SVM Classifiers with the Modified Particle Swarm Optimization and the SVM Ensembles. *International Journal of Advanced Computer Science and Applications (IJACSA).* 2016, vol. 7, no. 5, pp. 294–312.

15. Jain A. K., Ramaswami M. D. Classifier Design with Parzen Windows. *Pattern Recognition and Artificial Intelligence*. Elsevier Sci. Publ., 1988, vol. 7, pp. 211–228.

16. Li R., Wang H.-N., He H., Cui Y.-M., Zh.-L. Du. Support Vector Machine combined with K-Nearest Neighbors for Solar Flare Forecasting. *Chinese Journal of Astronomy and Astrophysics*, 2007, vol. 7, no. 3, pp. 441–447.

17. Mousa A., Yusof Y. An improved Chebyshev distance metric for clustering medical images. *AIP Conference Proceedings*, 2015, vol. 1691, issue 1, pp. 040020.

18. Pan Zh.-W., Xiang D. H., Xiao Q. W., Zhou D.-X. Parzen windows for multi-class classification. *Journal of Complexity*, 2008, October–December, vol. 24, issues 5–6, pp. 606–618.

19. Parzen E. On Estimation of 3 Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 1962, vol. 33, no. 3, pp. 1065–1076.

20. Vapnik V. *Statistical Learning Theory*. New York, John Wiley & Sons, inc., 1998, p. 740.

21. Wang H., Bell D. Extended k-Nearest Neighbours Based on Evidence Theory. *The Computer Journal*. 2004, vol. 47 (6), pp. 662–672.

22. Weinberger K. Q., Saul L. K. Distance Metric Learning for Large Margin Nearest Neighbor Classification. *Journal of Machine Learning Research*, 2009, vol. 10, pp. 207–244.

23. Yu L., Wang S., Lai K. K., Zhou L. *Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines*. Springer-Verlag Berlin Heidelberg Publ., 2008, p. 244.

24. Zambom A. Z., Dias R. A Review of Kernel Density Estimation with Applications to Econometrics. *International Econometric Review (IER), Econometric Research Association*, 2013, vol. 5 (1), pp. 20–42.

25. Zhang H., Berg A. C., Maire M., Malik J. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. *Proceedings – 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2126–2136.