УДК 004.855.5

КЛАССИФИКАЦИЯ ДАННЫХ НА ОСНОВЕ SVM-АЛГОРИТМА И АЛГОРИТМА *k*-БЛИЖАЙШИХ СОСЕДЕЙ

Л. А. Демидова, д.т.н., профессор кафедры ВПМ РГРТУ; liliya.demidova@rambler.ru
Ю. С. Соколова, старший преподаватель кафедры ВПМ РГРТУ; JuliaSokolova62@yandex.ru

Рассматривается задача классификации сложноорганизованных многомерных данных, характерная для различных социально-экономических, технических и других систем.

***Целью работы** является повышение точности классификации сложноорганизованных многомерных данных посредством разработки двухэтапного метода их классификации, основанного на совместном применении SVM- и kNN-классификаторов. На первом этапе метода классификации на основе исходного учебного набора данных U разрабатывается SVM-классификатор и определяется ширина Ω-области, содержащей все ошибочно классифицированные SVM-классификатором объекты, формирующие вместе с правильно классифицированными объектами, попавшими в Ω-область и соответствующими метками классов объектов из Ω-области, новый набор данных G. На втором этапе метода классификации ко всем объектам набора данных G из Ω-области применяется kNN-классификатор, разработанный на основе информации об объектах набора U\G. В случае улучшения качества классификации объектов, принадлежащих Ω-области, предлагаемый двухэтапный метод может быть рекомендован для классификации новых объектов. Значения параметров kNN-классификатора определяются экспериментально таким образом, чтобы обеспечить максимально возможную точность классификации объектов. Поскольку в формируемую вышеуказанным образом Ω-область могут попасть и верно классифицируемые объекты, то условием применимости предлагаемого метода является общее повышение качества классификации. Приведенные результаты экспериментальных исследований подтверждают эффективность применения предлагаемого метода в задаче классификации сложноорганизованных многомерных данных.*

***Ключевые слова**: SVM-классификатор, опорные векторы, тип функции ядра, параметры функции ядра, параметр регуляризации, kNN-классификатор, метод классификации.*

DOI: 10.21667/1995-4565-2017-62-4-119-132

Введение

В системах интеллектуального анализа данных (Data Mining, DM) особое место занимает проблема классификации, поскольку необходимость в её проведении возникает при решении широкого круга прикладных задач, связанных, например, с анализом кредитного риска, медицинской диагностикой, распознаванием рукописных символов, категоризацией текстов, извлечением информации, идентификацией изображений пешеходов, идентификацией изображений лиц, атрибуцией произведений искусства и т.п.

В настоящее время для разнообразных прикладных задач, использующих данные различных природы и объемов, разработаны десятки алгоритмов и методов классификации и их модификаций, среди которых наиболее известны, в частности, линейная и логистическая регрессии, байесовский классификатор, деревья решений, решающие правила, нейронные сети, алгоритм k -ближайших соседей (kNN-алгоритм, k Nearest Neighbors Algorithm), алгоритм опорных векторов (SVM-алгоритм, Support Vector Machine Algorithm) и т.п. [1-4].

Разработка того или иного классификатора предполагает выполнение процедур обучения и тестирования, при приемлемом качестве которых классификатор может быть применен для классификации новых объектов. Для оценки качества построенного классификатора могут использоваться различные общеизвестные показатели качества классификации, такие как: показатель F -меры, показатель чувствительности, показатель специфичности, показатель точности, показатель полноты и т.п. [4, 5]. Кроме того, может быть выполнен анализ ROC-кривой.

Следует отметить, что не существует универсальных алгоритмов и методов классификации. Более того, применение различных инструментов моделирования к одному и тому же набору объектов может привести к различным результатам. Это связано с тем, что в основу этих

инструментов заложены различные принципы моделирования, различаются и используемые в них метрики (меры расстояния), функции близости, критерии оптимальности, алгоритмы оптимизации, способы выбора начальных приближений, способы работы с разнотипными характеристиками и т.п. [1-10].

В последние годы для решения многих классификационных задач успешно применяется SVM-алгоритм, осуществляющий обучение по прецедентам («обучение с учителем») и входящий в группу граничных алгоритмов и методов классификации. SVM-алгоритм обеспечивает построение бинарного SVM-классификатора, реализуя с помощью специальной функции, называемой функцией ядра, перевод векторов характеристик классифицируемых объектов в пространство более высокой размерности и поиск в этом пространстве гиперплоскости с максимальным зазором, разделяющей объекты с разной классовой принадлежностью [1-4]. По обеим сторонам разделяющей гиперплоскости строятся две параллельные гиперплоскости, задающие границы классов и находящиеся на максимально возможном расстоянии друг от друга. Предполагается, что чем больше расстояние между этими параллельными гиперплоскостями, тем увереннее можно классифицировать объекты [1-4].

Несмотря на то, что способность к обобщению у SVM-алгоритма лучше, чем у других алгоритмов и методов классификации, существуют трудности в его применении, связанные с выбором типа функции ядра, значений параметров функции ядра и значения параметра регуляризации, влияющих на качество классификации данных.

Используемые при построении SVM-классификатора тип функции ядра, значения параметров функции ядра и значение параметра регуляризации предлагается определять с помощью модифицированного PSO-алгоритма [5, 11-13], позволяющего сократить временные затраты на поиск оптимальных значений параметров SVM-классификатора, что очень важно при классификации сложноорганизованных многомерных данных больших объемов. Значения параметров SVM-классификатора будем считать оптимальными, если достигнута высокая точность классификации: количество ошибок на обучающем и тестовом наборах минимально, причем количество ошибок обученного SVM-классификатора на объектах тестовой выборки не сильно отличается от количества ошибок на обучающей выборке (во избежание «переобучения» SVM-классификатора).

В большинстве случаев SVM-классификатор, построенный на основе модифицированного PSO-алгоритма, обеспечивает высокое качество классификации данных при приемлемых временных затратах [11-13]. При этом, как показывают результаты экспериментальных исследований, большинство ошибочно классифицированных объектов попадают внутрь полосы, разделяющей классы. В связи с этим целесообразна разработка методов, позволяющих повысить точность классификационных решений посредством уменьшения числа ошибок внутри разделяющей полосы.

Один из современных подходов к решению проблемы повышения точности классификационных решений предполагает ансамблирование тем или иным образом различных классификаторов с целью получения итогового классификационного решения более высокого качества [4, 13-19].

Как показывают результаты экспериментальных исследований, ансамбли классификаторов в случае их правильной конфигурации и настройки делают меньшее число ошибок классификации, чем каждый из участников ансамбля по отдельности. В связи с этим возникает необходимость в анализе результирующего классификационного решения, полученного путем применения к одному набору объектов нескольких алгоритмов и методов классификации. Очевидно, что ансамблирование вполне успешно может быть применено и в случае применения SVM-классификатора. При этом возможно как создание ансамбля, состоящего только из одних SVM-классификаторов [4, 13-15], так и сочетание в ансамбле наряду с SVM-классификатором какого-либо другого классификатора, принципиально отличного от SVM-классификатора по применяемому в нем инструментарию [20, 21].

Поскольку разработка SVM-классификатора связана с существенными временными затратами на определение оптимальных типа функции ядра, значений параметров функции ядра и значения параметра регуляризации, то одним из немаловажных требований, которые могут быть предъявлены к вводимому в ансамбль новому классификатору, наряду с требованием об обеспечении высокой точности классификации, является требование о незначительности временных затрат на разработку классификатора.

В качестве такого классификатора, в частности, может быть использован *k*NN-классификатор на основе *k*NN-алгоритма, который при определенных условиях, описываемых ниже, позволит повысить общую точность классификации данных при обеспечении незначительного увеличения временных затрат. *k*NN-классификатор является простейшим метрическим классификатором, основанным на оценке сходства некоторого объекта с другими k объектами – его

ближайшими соседями. При этом классифицируемый объект относится к тому классу, к которому принадлежит большинство из его k ближайших соседей-объектов [1, 2, 6, 7].

В настоящее время известен ряд подходов, реализующих совместное использование SVM- и kNN-классификаторов. Так, в [20] предлагается применять локальный SVM-классификатор для классификации объекта, ошибочно классифицированного kNN-классификатором, используя данные о ближайших соседях этого объекта при разработке SVM-классификатора. В [21] предлагается использовать при разработке kNN-классификатора, который должен уточнить классовую принадлежность объектов внутри разделяющей полосы, информацию об опорных векторах SVM-классификатора. Тем не менее, все предложенные подходы пока еще далеки от совершенства, а их эффективность очевидна лишь при решении задач классификации с определенным типом структуры данных. Очевидно, в частности, что разработка локальных SVM-классификаторов сопряжена с дополнительными временными затратами как на определение оптимального числа соседей классифицируемого объекта, так и на непосредственную разработку SVM-классификатора, а использование информации об опорных векторах SVM-классификатора при разработке kNN-классификатора требует предварительной тщательной проверки объективности их определения.

В данной работе предлагается двухэтапный метод классификации, основанный на совместном применении SVM- и kNN-алгоритмов и обеспечивающий повышение точности классификации сложноорганизованных многомерных данных. При этом на первом этапе метода классификации на основе исходного учебного набора данных об объектах будет разрабатываться SVM-классификатор и определяться ширина Ω-области, содержащей все объекты, ошибочно классифицированные SVM-классификатором. На втором этапе метода классификации на основе данных обо всех объектах, оказавшихся вне Ω-области, будет разрабатываться kNN-классификатор с целью улучшения качества классификации объектов, принадлежащих Ω-области. В случае улучшения качества классификации объектов, принадлежащих Ω-области, предлагаемый двухэтапный метод может быть применен для классификации новых объектов.

Разработка SVM-классификатора

При разработке SVM-классификатора используется учебный набор данных: $U = \{< z_1, y_1 >,...,< z_s, y_s >\}$, в котором каждый кортеж $< z_i, y_i >$ содержит информацию об объекте $z_i \in Z$ и число $y_i \in Y = \{-1;+1\}$, определяющее метку класса, к которому принадлежит объект z_i [1-5, 11-19]. Набор объектов Z представляет собой объединение набора объектов Z^-, метка класса которых принимает значение «−1», и набора объектов Z^+, метка класса которых принимает значение «+1», т.е. $Z = Z^- \cup Z^+$. Каждый объект $z_i \in Z$ представлен q-мерным вектором числовых характеристик $z_i = (z_i^1, z_i^2, \ldots, z_i^q)$ (нормированных значениями из отрезка [0, 1]), где z_i^l – числовое значение l-й характеристики для i-го объекта ($i = \overline{1,s}$, $l = \overline{1,q}$) [11-19].

Вышеуказанный учебный набор данных U многократно случайным образом разбивается на обучающую и тестовую выборки, состоящие соответственно из S и $s - S$ кортежей ($s > S$), с целью реализации многократного обучения и тестирования формируемых SVM-классификаторов с последующим определением лучшего классификатора в смысле обеспечения максимально возможной точности классификации. При этом для каждого SVM-классификатора определяются тип функции ядра $\kappa(z_i, z_\tau)$, значения параметров функции ядра и значение параметра регуляризации C ($C > 0$), позволяющего найти компромисс между максимизацией ширины полосы, разделяющей классы, и минимизацией суммарной ошибки.

При разработке SVM-классификатора с применением функции ядра $\kappa(z_i, z_\tau)$ определяется классифицирующая функция $F: Z \to Y$, устанавливающая для объекта $z_i \in Z$ его класс принадлежности $y_i \in Y = \{-1; +1\}$. В качестве функции ядра $\kappa(z_i, z_\tau)$ обычно используются следующие функции: линейная $\kappa(z_i, z_\tau) = z_i \bullet z_\tau$; полиномиальная $\kappa(z_i, z_\tau) = (z_i \bullet z_\tau + 1)^d$; радиальная базисная $\kappa(z_i, z_\tau) = exp(-(z_i - z_\tau) \bullet (z_i - z_\tau)/(2 \cdot \sigma^2))$; сигмоидная $\kappa(z_i, z_\tau) = th(k_2 + k_1 \cdot z_i \bullet z_\tau)$, где $z_i \bullet z_\tau$ – скалярное произведение векторов z_i и z_τ; d [$d \in N$ (по умолчанию $d = 3$)], σ [$\sigma > 0$ (по умолчанию $\sigma^2 = 1$)], k_2 [$k_2 < 0$ (по умолчанию $k_2 = -1$)] и k_1 [$k_1 > 0$ (по умолчанию $k_1 = 1$)] – некоторые параметры; th – гиперболический тангенс [1-5, 11-19].

В случае линейной разделимости классов в результате обучения SVM-классификатора строится гиперплоскость, разделяющая объекты из

Z на два класса: $w \bullet z + b = 0$, где w – вектор нормали к гиперплоскости, b – параметр, задающий смещение гиперплоскости относительно начала координат, $w \bullet z$ – скалярное произведение вектора w нормали к гиперплоскости и вектора характеристик некоторого объекта z [1–4]. Условие $-1 < w \bullet z + b < 1$ задает полосу, которая разделяет классы. Чем шире эта полоса, тем увереннее можно классифицировать объекты. Для максимизации ширины полосы $2/(w \bullet w)$ таким образом, чтобы внутрь нее не попал ни один объект из обучающей выборки, должна быть решена задача квадратичной оптимизации [1–4]:

$$\begin{cases} w \bullet w \to \min, \\ y_i \cdot (w \bullet z_i + b) \geq 1, \ i = \overline{1,S}. \end{cases}$$

В случае линейной неразделимости классов задача построения разделяющей гиперплоскости (с учетом теоремы Куна – Таккера) сводится к задаче квадратичного программирования, содержащей только двойственные переменные λ_i ($i = \overline{1,S}$) [1–4]:

$$\begin{cases} -L(\lambda) = \frac{1}{2} \cdot \sum_{i=1}^{S} \sum_{\tau=1}^{S} \lambda_i \cdot \lambda_\tau \cdot y_i \cdot y_\tau \cdot \kappa(z_i, z_\tau) - \sum_{i=1}^{S} \lambda_i \to \min\limits_{\lambda}, \\ \sum_{i=1}^{S} \lambda_i \cdot y_i = 0, \\ 0 \leq \lambda_i \leq C, \ i = \overline{1,S}. \end{cases}$$

В результате обучения SVM-классификатора определяются опорные векторы, являющиеся векторами характеристик тех объектов z_i из обучающей выборки, для которых значения соответствующих им двойственных переменных λ_i отличны от нуля ($\lambda_i \neq 0$) [1-4]. Опорные векторы находятся ближе всего к разделяющей гиперплоскости и несут всю информацию о разделении классов.

В результате обучения определяется классифицирующая функция, устанавливающая для произвольного объекта z его класс принадлежности с меткой «−1» или «+1» [1-4]:

$$F(z) = sign\left(\sum_{i=1}^{S} \lambda_i \cdot y_i \cdot \kappa(z_i, z) + b \right), \qquad (1)$$

где $b = w \bullet z_i - y_i$; $w = \sum_{i=1}^{S} \lambda_i \cdot y_i \cdot z_i$.

При этом суммирование в правиле (1) выполняется только по опорным векторам (то есть по векторам, для которых $\lambda_i \neq 0$).

Основная проблема, возникающая при разработке SVM-классификатора, связана с отсутствием рекомендаций по выбору типа функции ядра $\kappa(z_i, z_j)$, значений параметров функции ядра и значения параметра регуляризации C, при которых будет обеспечена высокая точность классификации объектов. Данная проблема может быть решена с применением тех или иных оптимизационных алгоритмов, например с использованием PSO-алгоритма, хорошо зарекомендовавшего себя при решении широкого спектра задач оптимизации.

PSO-алгоритм является алгоритмом случайно-направленного поиска. Данный алгоритм работает со случайно сгенерированной популяцией решений и выполняет расчет значений целевой функции, осуществляя в процессе эволюции поиск лучшего решения [22]. В частности, при разработке SVM-классификатора для выбора параметров, обеспечивающих высокую точность классификации объектов, может быть использован традиционный или модифицированный PSO-алгоритм [4, 11-13].

Поскольку большая часть ошибочно-классифицированных SVM-классификатором объектов располагается вблизи гиперплоскости, разделяющей классы, то еще один подход к повышению качества классификации заключается в применении дополнительного инструментария, способствующего повышению качества классификации для объектов, расположенных вблизи разделяющей гиперплоскости. В качестве такого инструментария может быть использован *k*NN-классификатор [1, 6, 7].

Разработка *k*NN-классификатора

При разработке *k*NN-классификатора на основе *k*NN-алгоритма, как и при разработке SVM-классификатора на основе SVM-алгоритма, используется учебный набор данных: $U = \{< z_1, y_1 >, ..., < z_s, y_s >\}$, который также случайным образом разбивается на обучающую и тестовую выборки, состоящие соответственно из S и $s - S$ кортежей ($s > S$) с целью реализации многократного обучения и тестирования формируемых в результате *k*NN-классификаторов с последующим определением лучшего классификатора в смысле обеспечения максимально возможной точности классификации. При этом для каждого *k*NN-классификатора определяется значение числа соседей k, при котором ошибка классификации минимальна. Класс принадлежности $y_i \in Y$ объекта $z_i \in Z$ определяется классом принадлежности большинства объектов из числа k ближайших соседей объекта $z_i \in Z$.

Реализация *k*NN-алгоритма для определения класса принадлежности произвольного объекта z при фиксированном числе k ближайших сосе-

дей предполагает выполнение следующей последовательности шагов.

1. Вычислить расстояние $d(z,z_i)$ от объекта z до каждого из объектов z_i, классовая принадлежность которых известна. Выполнить упорядочение вычисленных расстояний по возрастанию их значений.

2. Выбрать k объектов z_i (k ближайших соседей), наиболее близко расположенных к объекту z.

3. Выявить классовую принадлежность каждого из k ближайших соседей объекта z. Установить для произвольного объекта z в качестве его класса принадлежности класс, наиболее характерный для его k ближайших соседей.

Для оценки расстояния между объектами в kNN-алгоритме могут использоваться различные меры расстояния, такие как евклидова мера, манхэттенская мера, косинусная мера и др. [1].

При этом наиболее часто используется евклидова мера расстояния [1, 6]:

$$d(z_i,\ z)=(\sum_{l=1}^{q}(z_i^l-z^l)^2)^{1/2}, \qquad (2)$$

где q – число характеристик объектов z_i и z.

При реализации kNN-алгоритма могут применяться такие способы голосования, как простое невзвешенное голосование и взвешенное голосование [1, 6].

При использовании простого невзвешенного голосования расстояние от объекта z до каждого из k объектов – ближайших соседей z_i ($i=\overline{1,k}$) не играет роли: все k объектов – ближайших соседей z_i ($i=\overline{1,k}$) имеют равные права в определении класса объекта z. Каждый из k объектов – ближайших соседей z_i ($i=\overline{1,k}$) объекта z голосует за его отнесение к своему классу $y_{z_i,z}$. В результате реализации kNN-алгоритма объект z будет отнесен к тому классу, который наберет большее число голосов:

$$\alpha=arg\max_{y\in Y}\sum_{i=1}^{k}|\,y_{z_i,z}=y\,|. \qquad (3)$$

При использовании взвешенного голосования учитывается расстояние от объекта z до каждого из k объектов – ближайших соседей z_i ($i=\overline{1,k}$): чем меньше расстояние, тем более значимый вклад в оценку принадлежности объекта z к некоторому классу вносит голос объекта-соседа z_i ($i=\overline{1,k}$).

Оценка суммарного вклада голосов объектов-соседей z_i ($i=\overline{1,k}$) за принадлежность объекта z классу $y\in Y$ при взвешенном голосовании может быть рассчитана как [1]:

$$\alpha=\sum_{i=1}^{k}\frac{1}{d^2(z_i,\ z)}\cdot\rho_i, \qquad (4)$$

где $\rho_i=0$, если $y_{z_i,z}\neq y$ и $\rho_i=1$, если $y_{z_i,z}=y$.

Класс, которому соответствует наибольшее значение оценки (4), назначается рассматриваемому объекту z.

При невзвешенном голосовании расстояние между объектами z_i и z может быть вычислено на основе функции ядра [21]:

$$d^2(z_i,\ z)=\kappa(z_i,\ z_i)-2\cdot\kappa(z_i,\ z)+\kappa(z,\ z). \qquad (5)$$

Двухэтапный метод классификации

Как показывают результаты экспериментальных исследований, ни один классификатор данных не может быть признан несомненно лучшим по отношению к другим классификаторам, поскольку не позволяет обеспечить высокое качество классификации для любых произвольных наборов данных ввиду специфики применяемого при его разработке инструментария и соответственно ограниченности его возможностей.

Успешно применяемый в последние годы SVM-классификатор на подавляющем большинстве сложноорганизованных многомерных наборах данных обеспечивает приемлемое качество классификации [4, 11-14]. При этом анализ расположения объектов, ошибочно классифицированных SVM-классификатором, показал, что большинство из них попадает внутрь полосы, разделяющей классы и задаваемой условием $-1<\ w\bullet z+\ b\ <1$. Уточнение классификационного решения для объектов, оказавшихся внутри разделяющей полосы, предлагается осуществить с помощью двухэтапного метода классификации, основанного на совместном использовании SVM- и *k*NN-классификаторов.

Предлагаемый двухэтапный метод классификации данных может быть реализован следующим образом.

Этап 1. Разработка SVM-классификатора с определением возможности разработки *k*NN-классификатора.

1.1. Разработка SVM-классификаторов с последующим выбором лучшего из них в смысле обеспечения наиболее высокой точности классификации на учебном наборе данных U осуществляется на основе сформированных случайным образом обучающей и тестовой выборок. Заданные при разработке SVM-классификатора тип функции ядра, значения параметров ядра, значение параметра регуляризации определяют гиперплоскость, разделяющую объекты на два класса с метками «–1» и «+1». Оценка качества классификации данных осуществляется с при-

менением таких показателей качества, как показатель общей точности классификации ($Accur$), показатель специфичности (Sp), показатель чувствительности (Se), показатель F-меры, показатель AUC_{test}, рассчитанный по тестовой выборке, и др., а также число ошибок I и II рода: Er_I и Er_{II} и число ошибок на обучающей и тестовой выборках: Er_{train} и Er_{test}.

1.2. Определение областей Ω^- и Ω^+, содержащих все ошибочно классифицированные объекты, оказавшиеся в наборах Z^- и Z^+ соответственно; d_{Ω^-} – ширины области Ω^-; d_{Ω^+} – ширины области Ω^+; N_{Ω^-} – числа объектов в области Ω^-; N_{Ω^+} – числа объектов в области Ω^+. Формирование на основе областей Ω^- и Ω^+ итоговой Ω-области, включающей 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 $G = \{< z_1, y_1 >, ..., < z_{N_\Omega}, y_{N_\Omega} >\}$, в котором каждый кортеж $< z_i, y_i >$ содержит информацию об объекте z_i из Ω-области, и соответствующую объекту z_i метку класса $y_i \in Y = \{-1; +1\}$.

При этом возможно использование двух вариантов формирования Ω-области, в результате реализации которых будут получены:

– *асимметричная* относительно разделяющей гиперплоскости Ω-область: $\Omega = \Omega^- \cup \Omega^+$;

– *симметричная* относительно разделяющей гиперплоскости Ω-область, содержащая объекты, находящиеся относительно разделяющей гиперплоскости на расстоянии, не превышающем Δ: $\Delta = \max\{d_{\Omega^-}, d_{\Omega^+}\}$.

1.3. Формирование набора данных $W = U \setminus G$ удалением из учебного набора данных U кортежей набора данных G. Набор W будет состоять только из тех кортежей набора U, классовая принадлежность объектов для которых SVM-классификатором была определена правильно. Объекты этого набора впоследствии будут использованы для разработки *k*NN-классификатора. Поскольку в Ω-области помимо ошибочно классифицированных объектов может находиться и некоторое число объектов, классифицированных правильно, то возможно, что число кортежей в наборе W окажется существенно меньше, чем в U, и их будет недостаточно для последующей разработки *k*NN-классификатора.

Этап 2. Разработка *k*NN-классификатора.

2.1. Разработка на основе набора данных $W = U \setminus G$ *k*NN-классификаторов, устанавливающих классовую принадлежность всех объектов Ω- области при различных значениях числа k ближайших соседей из набора W с использованием различных способов голосования (3) и (4) и различных способов оценки близости между объектами [например, в соответствии с (2) и (5)]. Выбор лучшего *k*NN-классификатора в смысле обеспечения наиболее высокой точности классификации всех объектов Ω-области и фиксация значений параметров лучшего *k*NN-классификатора: варианта рассматриваемой Ω-области, используемого способа оценки близости между объектами, способа голосования и оптимального числа соседей.

2.2. Сравнение качества итоговой классификации данных с применением лучшего *k*NN-классификатора с качеством классификации, полученным после разработки SVM-классификатора, с целью выявления целесообразности применения сформированного таким образом классификатора для определения классовой принадлежности новых объектов, при этом целесообразность применения определяется улучшением показателей качества классификации объектов из набора Z.

В случае выявления целесообразности применения разработанного классификатора для классификации новых объектов она может быть выполнена в соответствии со следующей последовательностью шагов:

– разделить новые объекты на два класса с помощью разработанного SVM-классификатора;

– выделить из новых объектов те, которые оказались внутри Ω-области, сформированной в п. 1.2, и для этих объектов произвести уточнение классификационного решения с использованием *k*NN-классификатора (с определенными в п. 2.1 способом оценки близости между объектами, способом голосования и числом соседей, при которых улучшается качество классификации объектов из набора Z).

При реализации предлагаемого двухэтапного метода классификации разработка *k*NN-классификатора осуществляется при различных значениях числа k ближайших соседей. В случае бинарной классификации при использовании простого невзвешенного голосования логично использовать нечетные значения k во избежание ситуаций, когда за разные классы проголосовало одинаковое число соседей.

При разработке *k*NN-классификатора придется иметь дело уже не со всем исходным учебным набором данных U, а с набором существенно меньшей мощности, содержащим лишь информацию об объектах, попавших в Ω-область, а также об их классовой принадлежно-

сти. Использование нового дополнительного инструментария – *k*NN-классификатора, в котором используются принципы интеллектуального анализа данных, отличные от принципов, заложенных в SVM-классификатор, позволит в ряде случаев повысить общую точность классификации данных.

Ограничения на применимость предлагаемого двухэтапного метода классификации связаны с тем, что из-за большой ширины Ω-области или чрезмерной скученности объектов набора Z внутри Ω-области число кортежей в наборе $W = U \setminus G$ может оказаться недостаточным для последующей разработки *k*NN-классификатора.

Экспериментальные исследования

Экспериментальные исследования производились на основе ПЭВМ, работающей под 64-разрядной версией Windows 10 с оперативной памятью 8 Гб и двухядерным процессором Intel® Core™ i5-7200U с тактовой частотой каждого ядра 2,5 ГГц. В ходе исследований использовалась программная реализация SVM-алгоритма, предоставляемая системой инженерных и научных расчетов Matlab 7.12.0.635.

Для визуального представления результатов применения предложенного двухэтапного метода в задаче бинарной классификации был рассмотрен демонстрационный набор данных *Demo*, включающий 115 объектов с двумя характеристиками ($q = 2$). При этом класс принадлежности 95 объектов ($s = 95$), вошедших в учебный набор данных U, был заранее определен, а классовую принадлежность еще 20 объектов, вошедших в набор V, необходимо было установить.

На рисунке 1 представлено расположение объектов из демонстрационного набора данных *Demo* в пространстве D-2 с разбиением набора объектов Z на набор $Z^{\times}$, объекты которого помечены маркером «крестик» и принадлежат первому классу, и набор Z^{*}, объекты которого помечены маркером «звездочка» и принадлежат второму классу ($Z = Z^{\times} \cup Z^{*}$). Объекты набора V с неизвестной классовой принадлежностью помечены маркером «треугольник».

В ходе реализации первого этапа предлагаемого двухэтапного метода классификации при разработке SVM-классификатора с использованием радиальной базисной функции с параметрами, заданными по умолчанию ($\sigma = 1$ и $C = 1$), на сформированных случайным образом обучающей и тестовой выборках была построена кривая, разделяющая классы (рисунок 2). Мощность тестовой выборки составила 20 % от мощности учебного набора данных U. Объекты каждого класса, вошедшие в тестовую выборку, помечены маркерами меньшего размера, чем объекты, образующие обучающую выборку. Объекты, класс принадлежности которых необходимо определить (помеченные маркером «треугольник» на рисунке 1), на рисунке 2 не представлены. Определенные SVM-классификатором опорные векторы (в количестве 24 штук) дополнительно помечены маркером «кружок». Объекты, классовая принадлежность которых SVM-классификатором определена неверно, обозначены номерами 1 – 4.

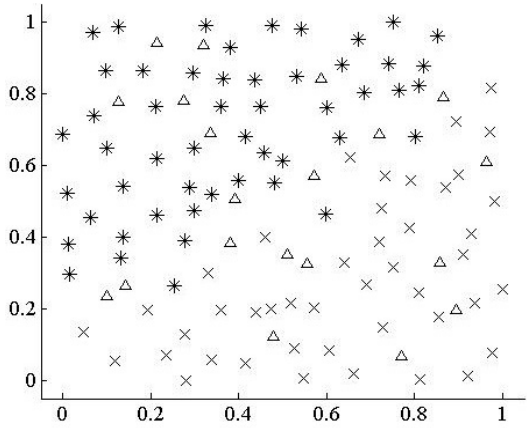

Рисунок 1 – Представление демонстрационного набора данных *Demo* в пространстве D-2

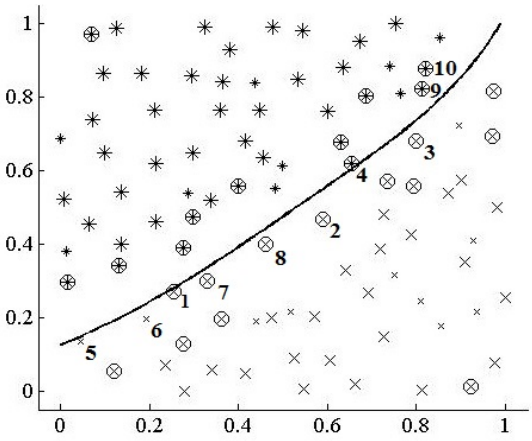

Рисунок 2 – Результаты разделения демонстрационного набора данных *Demo* на классы

Как видно из рисунка 2, три объекта из набора Z^{*}, а именно объекты с номерами 1, 2 и 3, ошибочно отнесены SVM-классификатором в набор $Z^{\times}$, сопоставленный первому классу, а объект с номером 4, принадлежащий набору $Z^{\times}$, ошибочно отнесен в набор Z^{*}, сопоставленный второму классу.

Таким образом, SVM-классификатор допустил 4 ошибки на обучающей выборке, а при классификации объектов, вошедших в тестовую выборку, SVM-классификатор не допустил ошибок.

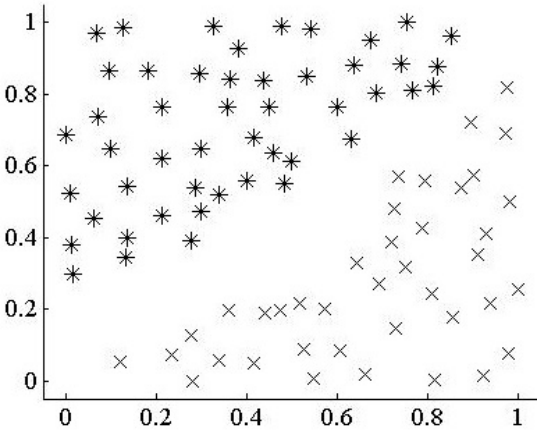

Рисунок 3 – Учебный набор данных без объектов Ω-области

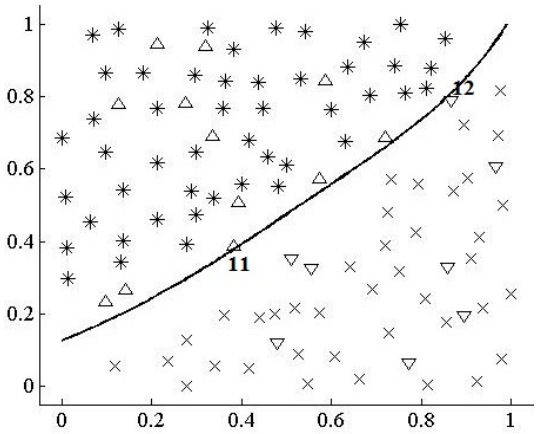

Рисунок 4 – Результат SVM-классификации (без Ω-области)

При этом все 4 ошибочно классифицированных объекта определены SVM-классификатором в качестве опорных векторов.

Оценка качества классификации показала, что значение показателя *F*-меры равно 96 %.

Перед реализацией второго этапа предлагаемого двухэтапного метода классификации предварительно для каждого класса были определены: области Ω^* и $\Omega^\times$, содержащие все ошибочно классифицированные объекты, оказавшиеся в наборах Z^* и $Z^\times$ соответственно; d_{Ω^*} – ширина области Ω^*; $d_{\Omega^\times}$ – ширина области $\Omega^\times$; N_{Ω^*} – число объектов в области Ω^*; $N_{\Omega^\times}$ – число объектов в области $\Omega^\times$, а также результирующая Ω-область, содержащая все ошибочно классифицированные объекты (при этом в Ω-области может находиться и некоторое число объектов, классифицированных правильно).

Необходимо отметить, что для демонстрационного набора данных *Demo* принятые в теоретической части обозначения для наборов Z^- и Z^+, областей Ω^- и Ω^+, ширины областей d_{Ω^-} и d_{Ω^+}, числа объектов N_{Ω^-} и N_{Ω^+} заменены в целях лучшей визуализации на обозначения Z^* и $Z^\times$, Ω^* и $\Omega^\times$, d_{Ω^*} и $d_{\Omega^\times}$, N_{Ω^*} и $N_{\Omega^\times}$ соответственно.

Уточнение результатов классификации для объектов, вошедших в Ω-область, содержащую все ошибочно классифицированные объекты, выполнялось с использованием *k*NN-алгоритма, при этом рассматривались два варианта Ω-области:

– *асимметричная* относительно разделяющей кривой Ω-область: $\Omega = \Omega^* \cup \Omega^\times$;

– *симметричная* относительно разделяющей кривой Ω-область, содержащая объекты, находящиеся относительно разделяющей кривой на расстоянии, не превышающем Δ: $\Delta = max\{d_{\Omega^*},\ d_{\Omega^\times}\}$.

Вариант 1. При использовании асимметричной Ω-области на основе разработанного SVM-классификатора было получено, что $d_{\Omega^*} = 0{,}079$, $N_{\Omega^*} = 1$, $d_{\Omega^\times} = 0{,}503$, $N_{\Omega^\times} = 7$. Внутрь Ω-области попало 8 объектов (это объекты с номерами 1 – 8 на рисунке 2).

В результате удаления из учебного набора U кортежей набора G, содержащих информацию об объектах, оказавшихся в Ω-области, и соответствующих им метках класса, был получен набор данных $W = U \setminus G$, состоящий только из кортежей, для объектов которых классовая принадлежность SVM-классификатором была определена правильно (рисунок 3).

На основе информации о классовой принадлежности объектов набора данных $W = U \setminus G$ для объектов Ω-области была выполнена разработка *k*NN-классификатора при числе соседей, изменяющемся от 1 до 51 (с шагом 2), и различными способами оценки близости объектов. При оценке качества *k*NN-классификации выполнялось сравнение меток классов объектов Ω-области, определенных с помощью *k*NN-классификатора, и исходных меток классов объектов Ω-области, зафиксированных в учебном наборе U. В качестве лучшего был определен *k*NN-классификатор с числом соседей, равным 7. При этом оценка близости между объектом и его соседями осуществлялась с помощью евклидовой меры при невзвешенном голосовании. В итоге число ошибок классификации объектов в Ω-области удалось сократить с 4 до 2, повысив значение показателя *F*-меры до 98 % (что на 2 % выше значения показателя *F*-меры SVM-классификатора).

Вариант 2. При использовании симметричной Ω-области на основе разработанного SVM-

классификатора было получено, что $d_\Omega = 1{,}005$, $N_\Omega = 10$. Внутрь Ω-области попало 10 объектов (это объекты с номерами 1 – 10 на рисунке 2). В результате удаления из учебного набора U кортежей набора G, содержащих информацию об объектах, оказавшихся в Ω-области, и соответствующих им метках класса, был получен набор данных $W = U \setminus G$, состоящий только из кортежей, для объектов которых классовая принадлежность SVM-классификатором была определена правильно. На основе этого набора данных, как и в варианте 1, была выполнена разработка *k*NN-классификатора при числе соседей, изменяющемся от 1 до 51 (с шагом 2), и различными способами оценки близости объектов. При оценке качества *k*NN-классификации выполнялось сравнение меток классов объектов Ω-области, определенных с помощью *k*NN-классификатора, и исходных меток классов объектов Ω-области, зафиксированных в учебном наборе U. В качестве лучшего был определен *k*NN-классификатор с числом соседей, равным 3. При этом степень близости между объектом и его соседями определялась с помощью евклидовой метрики при невзвешенном голосовании. В итоге число ошибок классификации объектов в Ω-области удалось сократить с 4 до 3, повысив значение показателя *F*-меры до 97,03 % (что на 1,03 % выше значения показателя *F*-меры SVM-классификатора).

Поскольку в рассматриваемом случае при совместном применении SVM- и *k*NN-классификаторов произошло улучшение качества классификации объектов, принадлежащих Ω-области, предлагаемый двухэтапный метод был применен для классификации новых объектов (помеченных на рисунке 1 маркером «треугольник»). При этом при уточнении результатов классификации с использованием *k*NN-алгоритма для объектов, вошедших в Ω-область, качество классификации было улучшено при каждом варианте Ω-области (то есть и для симметричной, и для асимметричной Ω-области). Однако, поскольку для учебного набора данных U большее улучшение было достигнуто при рассмотрении асимметричной Ω-области, то и при классификации новых объектов, принадлежащих Ω-области, рассматривался ее асимметричный вариант.

В результате применения разработанного SVM-классификатора к новым объектам, они были разделены на наборы Z^* и $Z^\times$, соответствующие классам с метками «звездочка» и «крестик». На рисунке 4 новые объекты помечены маркером «треугольник». Объекты из обучающей и тестовой выборок, попавшие в асимметричную Ω-область, на рисунке 4 не показаны. Из 20 новых объектов два объекта попали в асимметричную Ω-область: на рисунке 4 они обозначены номерами 11 и 12, при этом SVM-классификатор определил объект с номером 11 в набор Z^*, соответствующий классу с меткой «звездочка», а объект с номером 12 – в набор $Z^\times$, соответствующий классу с меткой «крестик».

Для новых объектов, оказавшихся внутри асимметричной Ω-области (то есть для объектов с номерами 11 и 12), результаты классификации уточнялись с помощью *k*NN-классификатора. Значения параметров *k*NN-классификатора, обеспечивающие максимально возможную точность классификации объектов, были определены экспериментально: было установлено, что лучшее качество классификации обеспечивается при рассмотрении асимметричной Ω-области, в которой степень близости между объектами определяется с помощью евклидовой метрики при невзвешенном голосовании при оптимальном числе соседей, равном 7.

Применение *k*NN-классификатора к объекту с номером 11 не изменило его классовой принадлежности, а объект с номером 12 был отнесен *k*NN-классификатором к набору Z^*, соответствующему классу с меткой «звездочка», т.е. *k*NN-классификатор изменил результат SVM-классификации.

Целесообразность использования предлагаемого двухэтапного метода классификации была подтверждена и на реальных наборах данных, взятых из проекта Statlog и репозитория задач машинного обучения UCI Machine Learning Repository, традиционно используемых для тестирования разрабатываемых алгоритмов машинного обучения. 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/), кредитного скоринга (Firms и German, источники [4] и http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/) и обработки сигналов (Ionosphere, источник https://archive.ics.uci.edu/ml/machine-learning-databases/ionosphere/) (таблица 1). На всех наборах данных имеет место случай бинарной классификации. В столбце 1 таблицы 1 содержится название учебного набора данных, количество объектов и число характеристик у каждого объекта классификации.

Для каждого учебного набора в таблице 1 представлены значения параметров классифика-

торов и значения показателей качества классификации для двух разработанных на сформированных случайным образом обучающей и тестовой выборках SVM-классификаторов с использованием функции ядра, указанной в столбце 2 таблицы 1 (для радиальной базисной функции-ядра указывается сокращение *rbf*, для полиномиальной – *poly*). При разработке SVM-классификаторов использовались значения параметров ядра и значение параметра регуляризации, заданные по умолчанию ($\sigma = 1$, $d = 3$ и $C = 1$). Мощность тестовой выборки составляет 20 % от мощности учебного набора данных U.

Число определенных в результате обучения SVM-классификатора опорных векторов содержится в столбце 3 таблицы 1.

Столбец 7 таблицы 1 показывает, какой тип классификатора использовался для соответствующего набора данных: запись *SVM* означает, что разрабатывался SVM-классификатор на основе типа ядра из столбца 2; запись *+aSim* означает, что после разработки SVM-классификатора производилось уточнение результатов классификации с применением *k*NN-классификатором для объектов, попавших в асимметричную относительно разделяющей гиперплоскости Ω-область (вариант 1); запись *+Sim* означает, что после разработки SVM-классификатора производилось уточнение результатов классификации с применением *k*NN-классификатора для объектов, попавших в симметричную относительно разделяющей гиперплоскости Ω-область (вариант 2). Значение ширины Ω-области (d_Ω) и число объектов, оказавшихся внутри нее (N_Ω), приведены в столбцах 8 и 9 соответственно.

Таблица 1 – Значения параметров классификаторов и значения показателей качества классификации

Набор данных	SVM-классификация					Тип классификации	*k*NN-классификация								Оценки качества классификации						
	Функция ядра	Число опорных векторов	$AUCt_{est}$	Число ошибок			d_Ω	N_Ω	способ голосования						*F-мера*, %	*Accur*, %	*Se*, %	*Sp*, %	Число ошибок		
									невзвешенное		взвешенное		с функцией ядра								
				Er_{train}	Er_{test}				Число соседей	Число ошибок	Число соседей	Число ошибок	Число соседей	Число ошибок					Er_I	Er_{II}	Всего
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Firms (60 × 11)	rbf	47	0,9167	1 из 48	3 из 12	SVM	-	-	-	-	-	-	-	-	93,55	93,33	96,67	90,00	1	3	4
						+aSim	0,185	5	**41**	**0**	13	1	**41**	**0**	**100**	**100**	**100**	**100**	**0**	**0**	**0**
						+Sim	0,253	6	11	1	5	2	11	1	98,31	98,33	96,67	100	1	0	1
	poly	31	0,8438	0 из 48	4 из 12	SVM	-	-	-	-	-	-	-	-	93,10	93,33	90,00	96,67	3	1	4
						+aSim	0,915	6	**3**	**2**	7	3	1	3	**96,55**	**96,67**	**93,33**	**100**	**2**	**0**	**2**
						+Sim	1,078	8	9	2	11	2	3	2	96,55	96,67	93,33	100	2	0	2
WDBC (569 × 30)	rbf	425	0,9887	0 из 456	11 из 113	SVM	-	-	-	-	-	-	-	-	98,44	98,07	97,20	99,53	10	1	11
						+aSim	0,142	36	7	3	7	3	7	3	99,58	99,47	99,44	99,53	2	1	3
						+Sim	0,152	36	**13**	**1**	**13**	**1**	**13**	**1**	**99,86**	**99,82**	**100**	**99,53**	**0**	**1**	**1**
	rbf	425	0,9949	0 из 456	8 из 113	SVM	-	-	-	-	-	-	-	-	98,87	98,59	97,76	100	8	0	8
						+aSim	0,035	23	**5**	**0**	**5**	**0**	**5**	**0**	**100**	**100**	**100**	**100**	**0**	**0**	**0**
						+Sim	0,070	27	11	0	11	0	11	0	100	100	100	100	0	0	0
German (1000 × 24)	rbf	798	0,7473	0 из 800	50 из 200	SVM	-	-	-	-	-	-	-	-	96,54	95,00	99,57	84,33	3	47	50
						+aSim	0,921	208	29	45	45	43	29	45	96,96	95,70	97,86	90,67	15	28	43
						+Sim	1,284	209	45	43	**43**	**42**	45	43	**97,04**	**95,80**	**98,29**	**90,00**	**12**	**30**	**42**
	rbf	799	0,7552	0 из 800	53 из 200	SVM	-	-	-	-	-	-	-	-	96,33	94,70	99,43	83,67	4	49	53
						+aSim	1,186	219	**39**	**41**	43	45	**39**	**41**	**97,11**	**95,90**	**98,29**	**90,33**	**12**	**29**	**41**
						+Sim	1,796	225	35	47	47	48	35	47	96,72	95,30	99,00	86,67	7	40	47
Heart (270 × 13)	rbf	211	0,8962	1 из 216	9 из 54	SVM	-	-	-	-	-	-	-	-	96,67	96,30	96,67	95,83	5	5	10
						+aSim	0,542	35	**37**	**5**	25	6	**37**	**5**	**98,34**	**98,15**	**98,67**	**97,50**	**2**	**3**	**5**
						+Sim	0,780	43	17	6	31	7	17	6	98,00	97,78	98,00	97,50	3	3	6
	rbf	208	0,8389	1 из 216	12 из 54	SVM	-	-	-	-	-	-	-	-	95,68	95,19	96,00	94,17	6	7	13
						+aSim	0,504	40	**25**	**6**	25	8	**25**	**6**	**98,00**	**97,78**	**98,00**	**97,50**	**3**	**3**	**6**
						+Sim	0,601	43	27	8	47	8	27	8	97,32	97,04	96,67	97,50	5	3	8
Ionosphere (351 × 34)	rbf	217	0,9983	1 из 281	3 из 70	SVM	-	-	-	-	-	-	-	-	98,44	98,86	100	98,22	0	4	4
						+aSim	0,224	6	**9**	**2**	11	2	**9**	**2**	**99,21**	**99,43**	**99,21**	**99,56**	**1**	**1**	**2**
						+Sim	0,448	14	**9**	**2**	11	2	**9**	**2**	**99,21**	**99,43**	**99,21**	**99,56**	**1**	**1**	**2**
	rbf	220	0,9925	1 из 281	5 из 70	SVM	-	-	-	-	-	-	-	-	97,67	98,29	100	97,33	0	6	6
						+aSim	0,207	9	**11**	**3**	13	3	**11**	**3**	**98,80**	**99,15**	**98,41**	**99,56**	**2**	**1**	**3**
						+Sim	0,414	16	**11**	**3**	13	3	**11**	**3**	**98,80**	**99,15**	**98,41**	**99,56**	**2**	**1**	**3**

Для оценки качества каждого типа классификации (*SVM*, *+aSim*, *+Sim*) использовались общеизвестные в машинном обучении показатели качества классификации, такие как: показатель F-меры, показатель общей точности классификации ($Accur$), показатель чувствительности (Se), показатель специфичности (Sp), число ошибок I и II рода (Er_I и Er_{II}), число ошибок на обучающей и тестовой выборках (Er_{train} и Er_{test}), общее число ошибок, показатель AUC_{test}, рассчитанный по тестовой выборке. В столбцах 16–21 таблицы 1 представлены значения показателя F-меры, показателя общей точности классификации, показателя чувствительности, показателя специфичности, а также число ошибок I и II рода; в столбцах 4–6 приведены значение показателя AUC_{test}, число ошибок на обучающей и тестовой выборках; в столбце 22 представлено общее число ошибок, допущенных классификатором.

После разработки SVM-классификатора была выполнена оценка числа объектов, попавших внутрь разделительной полосы, то есть полосы, в которой выполняется условие $-1 < w \bullet z + b < 1$. Для рассматриваемых учебных наборов во всех случаях, представленных в таблице 1, все 100 % ошибочно классифицированных объектов оказались внутри разделительной полосы. Кроме того, для всех наборов анализировалась возможность применения *k*NN-классификатора: оценивалась ширина Ω-области, в которую попадают все ошибочно классифицированные SVM-класси-фикатором объекты из учебного набора данных, и количество объектов, оказавшихся внутри нее. Во всех экспериментах, информация по которым представлена в таблице 1, оказалось, что Ω-область расположена внутри разделительной полосы.

В результате удаления из каждого учебного набора U (таблица 1) кортежей, соответствующих объектам Ω-области, был получен набор данных $W = U \setminus G$, число кортежей в котором меньше числа кортежей в соответствующем учебном наборе данных.

Далее для объектов Ω-области производилось уточнение результатов классификации с помощью *k*NN-классификатора на основе кортежей набора $W = U \setminus G$ при различном числе соседей (изменяющемся от 1 до 51) с использованием различных способов голосования (а именно взвешенного и невзвешенного голосования) и различных способов оценки близости между объектами. Для каждого способа голосования определялось оптимальное число соседей (столбцы 10, 12, 14 таблицы 1), при которых число ошибок классификации минимально (столбцы 11, 13, 15 таблицы 1).

После применения *k*NN-классификатора с числом соседей, способом оценки близости и способом голосования, обеспечивающими минимальное число ошибок классификации, при разных вариантах (асимметричном и симметричном) расположения Ω-области относительно разделяющей гиперплоскости, качество итоговой классификации данных было вновь оценено посредством расчета значений различных показателей. Кроме того, было выполнено сравнение новых значений показателей качества со значениями этих же показателей, полученных на основе разработанного SVM-классификатора. Ячейки таблицы 1, соответствующие лучшему варианту *k*NN-классификатора, обеспечивающему минимальное число ошибок классификации, выделены жирным шрифтом.

Из таблицы 1 видно, что в большинстве случаев максимальное уточнение результатов классификации достигается при использовании асимметричной Ω-области, но в ряде экспериментов двухэтапный метод классификации дает одинаковые результаты при использовании и асимметричной Ω-области, и симметричной Ω-области, а иногда максимальное уточнение результатов классификации достигается при использовании симметричной Ω-области. Также из таблицы 1 видно, что нельзя однозначно определить способ голосования и способ оценки близости, при которых бы всегда достигалось лучшее уточнение результатов классификации при применении *k*NN-классификатора, поскольку в ряде экспериментов лучшее решение имеет место при невзвешенном способе голосования, а иногда – при взвешенном.

В таблице 2 приведены значения показателей *F*-меры, показателя точности ($Accur$), суммарное число ошибок I и II рода (Er_I+Er_{II}) до и после применения *k*NN-классификатора, обеспечившего лучшее уточнение результатов классификации, для классификации объектов, попавших в Ω-область, символом δ обозначена разница между соответствующими значениями показателей SVM-классификации и нового классификатора, полученного при реализации предложенного двухэтапного метода классификации.

По данным таблицы 2 видно, что использование предлагаемого двухэтапного метода классификации позволило повысить в среднем значение показателя *F*-меры на 1,96 %, а значение показателя точности классификации на 2,1 % по сравнению с соответствующими значениями, полученными только при использовании SVM-классификатора, причем максимальное повыше-

ние показателя *F*-меры составило 6,45 %, а показателя точности классификации – 6,67 %.

Таблица 2 – Сравнение классификаторов

Набор данных	*F-мера*, %			*Accur*, %			Er_I+Er_{II}		
	SVM	SVM + *k*NN	δ	SVM	SVM + *k*NN	δ	SVM	SVM + *k*NN	δ
Firms	93,55	100	6,45	93,33	100	6,67	4	0	-4
	93,10	96,55	3,45	93,33	96,67	3,34	4	2	-2
WDBC	98,44	99,86	1,42	98,07	99,82	1,75	11	1	-10
	98,87	100	1,13	98,59	100	1,41	8	0	-8
German	96,54	97,04	0,5	95,00	95,80	0,8	50	42	-8
	96,33	97,11	0,78	94,70	95,90	1,2	53	41	-12
Heart	96,67	98,34	1,67	96,30	98,15	1,85	10	5	-5
	95,68	98,00	2,32	95,19	97,78	2,59	13	6	-7
Ionosphere	98,44	99,21	0,77	98,86	99,43	0,57	4	2	-2
	97,67	98,80	1,13	98,29	99,15	0,86	6	3	-3
Среднее δ			1,96			2,10			-6

Заключение

По результатам проведенных экспериментальных исследований можно сделать вывод о том, что использование предложенного двухэтапного метода повышает качество результатов классификации, так как применение *k*NN-классификатора к объектам, расположенным вблизи гиперплоскости, разделяющей классы и определенной SVM-классификатором, уменьшает число ошибочно классифицированных объектов. Предлагаемый двухэтапный метод классификации позволяет принимать высокоточные решения по классификации сложноорганизованных многомерных данных.

В ходе дальнейших исследований планируется рассмотреть модификации *k*NN-алгоритма, реализующие взвешенные варианты учета ближайших соседей с применением различных весовых функций, оценивающих степень важности i-го соседа для классификации объекта z, в частности алгоритм k экспоненциально взвешенных ближайших соседей, алгоритм парзеновского окна фиксированной (или переменной) ширины, алгоритм потенциальных функций, а также алгоритмы быстрого поиска ближайших соседей.

Библиографический список

1. **Воронцов К.В.** Математические методы обучения по прецедентам (теория обучения машин). 141 с. [Электронный ресурс]. URL: www.MachineLearning.ru (дата обращения: 27.12.2016).

2. **Вьюгин В.В**. Элементы математической теории машинного обучения: учеб. пособие. М.: МФТИ, 2010. 252 с.

3. **Vapnik V.** Statistical Learning Theory. New York: John Wiley & Sons. 1998. 732 p.

4. **Lean Yu, Shouyang Wang, Kin Keung Lai, Ligang Zhou**. Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines. Springer-Verlag Berlin Heidelberg, 2008. 244 p.

5. **Демидова Л.А., Соколова Ю.С.** Аспекты применения алгоритма роя частиц в задаче разработки SVM-классификатора // Вестник Рязанского государственного радиотехнического университета. 2015. № 53. С. 84-92.

6. **Сегаран Т.** Программируем коллективный разум. С П б .: Символ-Плюс, 2008. 368 с.

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

8. **Bezdek J.C., Keller J., Krisnapuram R., Pal N.R.** Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Springer Science+ Business Media. 2005. 785 p.

9. **Демидова Л.А., Коняева Е.И.** Кластеризация объектов с использованием FCM-алгоритма на основе нечетких множеств второго типа и генетического алгоритма // Вестник Рязанского государственного радиотехнического университета. 2008. № 26. С. 46-54.

10. **Демидова Л.А., Тишкин Р.В., Юдаков А.А.** Разработка ансамбля алгоритмов кластеризации на основе матриц подобия меток кластеров и алгоритма спектральной факторизации // Вестник Рязанского государственного радиотехнического университета. 2013. № 4-1 (46). С. 9-17.

11. **Demidova L., Nikulchev E., Sokolova Yu**. The SVM Classifier Based on the Modified Particle Swarm Optimization // International Journal of Advanced Computer Science and Applications (IJACSA). 2016. Vol. 7. no. 2. P. 16-24.

12. **Demidova L., Sokolova Yu**. Modification Of Particle Swarm Algorithm For The Problem Of The SVM Classifier Development // В сборнике: 2015 International Conference «Stability and Control Processes» in Memory of V.I. Zubov (SCP). 2015. P. 623-627.

13. **Демидова Л.А., Никульчев Е.В., Соколова Ю.С.** Классификация больших данных: использование SVM-ансамблей и SVM-классификаторов с модифицированным роевым алгоритмом // Cloud of Science. 2016. Vol. 3. № 1. P. 5-42.

14. **Демидова Л.А., Соколова Ю.С.** Разработка ансамбля SVM-классификаторов с использованием декорреляционного алгоритма максимизации // Информатика и системы управления. 2016. № 1 (47). С. 95-105.

15. **Demidova L., Sokolova Yu.** Development of the SVM Classifier Ensemble for the Classification Accuracy Increase // 6th Seminar on Industrial Control Systems: Analysis, Modeling and Computation (ITM Web of Conferences). 2016. Vol. 6.

16. **Демидова Л.А., Соколова Ю.С.** Использование SVM-алгоритма для уточнения решения задачи классификации объектов с применением алгоритмов кластеризации // Вестник Рязанского государственного радиотехнического университета. 2015. № 1 (51). С. 103-113.

17. **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. P. 446-457.

18. **Demidova L., Sokolova Yu**. SVM-Classifier Development With Use Of Fuzzy Clustering Algorithms' Ensemble On The Base Of Clusters' Tags' Vectors' Simi-

larity Matrixes // В сборнике: 16th International Symposium on Advanced Intelligent Systems 2015. P. 889-906.

19. **Demidova L., Sokolova Yu**. Training Set Forming For SVM Algorithm With Use Of The Fuzzy Clustering Algorithms Ensemble On Base Of Cluster Tags Vectors Similarity Matrices // В сборнике: 2015 International Conference «Stability and Control Processes» in Memory of V.I. Zubov (SCP). 2015. P. 619-622.

20. **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. Vol. 2. 2006. P. 2126-2136.

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

22. **Jun Sun, Choi-Hong Lai, Xiao-Jun Wu**. Particle Swarm Optimisation: Classical and Quantum Perspectives. CRC Press, 2011. 419 p.

UDC 004.855.5

TWO-STAGE DATA CLASSIFICATION METHOD BASED ON SVM-ALGORITHM AND THE *k* NEAREST NEIGHBORS ALGORITHM

L. A. Demidova, PhD (technical sciences), full professor, RSREU, Ryazan; liliya.demidova@rambler.ru
Yu. S. Sokolova, senior teacher, RSREU, Ryazan; JuliaSokolova62@yandex.ru

The classification problem of elaborate multidimensional data which is inherent in various socio-economic, technical and other systems has been considered. ***The aim is*** *the classification accuracy increase of elaborate multidimensional data by means of development of two-stage classification method based on the combined use of SVM and kNN classifiers. At the first stage of the classification method SVM classifier on the base of initial learning dataset U is developed and the width of Ω-area containing all objects classified erroneously by SVM classifier. Objects classified erroneously together with correctly classified objects which are also located in the Ω-area and the corresponding classes tags of objects from Ω-area form the new G dataset. At the second stage of the classification method kNN classifier developed on the base of information about the objects of U\G set is applied to all objects of G data set from Ω-area. In case of improvement of the classification quality of objects belonging to Ω-area, the offered two-stage method can be recommended for classification of new objects. The parameters values of kNN classifier are defined experimentally to provide the greatest possible classification accuracy of objects. As the correctly classified objects can also get to Ω-area created in the above-stated way, the condition of applicability of the offered method is general improvement of classification quality. The given results of experimental studies confirm the efficiency of the offered method application in the classification problem of elaborate multidimensional data.*

Key words: *SVM classifier, support vectors, kernel function type, kernel function parameters, regularization parameter, kNN classifier, classification method.*

DOI: 10.21667/1995-4565-2017-62-4-119-132

References

1. **Voroncov K. V.** Matematicheskie metody obuchenija po precedentam (teorija obuchenija mashin) (Mathematical methods of training in precedents (theory of machine training), 141 p., [Jelektronnyj resurs]. URL: www.MachineLearning.ru (data obrashhenija: 27.12.2016) (in Russian).

2. **V'jugin V. V.** Jelementy matematicheskoj teorii mashinnogo obuchenija: ucheb. posobie (Elements of the mathematical theory of machine training), M.: MFTI, 2010, 252 p. (in Russian).

3. **Vapnik V.** Statistical Learning Theory, New York: John Wiley & Sons, 1998, 732 p.

4. **Lean Yu, Shouyang Wang, Kin Keung Lai, Ligang Zhou**. Bio-Inspired Credit Risk Analysis. Computational Intelligence with Support Vector Machines. Springer-Verlag Berlin Heidelberg, 2008, 244 p.

5. **Demidova L. A., Sokolova Ju. S.** Aspekty primenenija algoritma roja chastic v zadache razrabotki SVM-klassifikatora (Aspects of application of the particle swarm algorithm in the problem of the SVM classifier development). Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta. 2015. no. 53. pp. 84-92 (in Russian).

6. **Segaran T.** Programmiruem kollektivnyj razum. SPb: Simvol-Pljus (Programming of the collective intelligence), 2008, 368 p. (in Russian).

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

8. **Bezdek J. C., Keller J., Krisnapuram R., Pal N. R.** Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Springer Science+ Business Media, 2005, 785 p.

9. **Demidova L.A., Konjaeva E.I.** Klasterizacija ob'ektov s ispol'zovaniem FCM-algoritma na osnove ne-

chetkih mnozhestv vtorogo tipa i geneticheskogo algoritma (Clustering of objects with the use of the FCM algorithm on the base of the second type fuzzy sets and the genetic algorithm), Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta. 2008. no. 26. pp. 46-54 (in Russian).

10. **Demidova L. A., Tishkin R. V., Judakov A. A.** Razrabotka ansamblja algoritmov klasterizacii na osnove matric podobija metok klasterov i algoritma spektral'noj faktorizacii (Development of the clustering algorithms ensemble on the base of the clusters tags similarity matrixes and the spectral factorization algorithm), Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta, 2013, no. 4-1 (46), pp. 9-17 (in Russian).

11. **Demidova L., Nikulchev E., Sokolova Yu**. The SVM Classifier Based on the Modified Particle Swarm Optimization // International Journal of Advanced Computer Science and Applications (IJACSA), 2016, vol. 7, no. 2, pp. 16-24.

12. **Demidova L., Sokolova Yu**. Modification Of Particle Swarm Algorithm For The Problem Of The SVM Classifier Development, 2015 International Conference «Stability and Control Processes» in Memory of V.I. Zubov (SCP), 2015, pp. 623-627.

13. **Demidova L. A., Nikul'chev E. V., Sokolova Ju.S.** Klassifikacija bol'shih dannyh: ispol'zovanie SVM-ansamblej i SVM-klassifikatorov s modificirovannym roevym algoritmom (Big data classification: use of the SVM ensembles and SVM classifiers with the modified swarm algorithm), Cloud of Science, 2016, vol. 3, no. 1, pp. 5-42 (in Russian).

14. **Demidova L.A., Sokolova Ju.S.** Razrabotka ansamblja SVM-klassifikatorov s ispol'zovaniem dekorreljacionnogo algoritma maksimizacii (Development of the SVM classifiers ensemble with the use of the maximizing decorrelation algorithm), Informatika i sistemy upravlenija, 2016, no. 1 (47), pp. 95-105 (in Russian).

15. **Demidova L., Sokolova Yu.** Development of the SVM Classifier Ensemble for the Classification Accuracy Increase, 6th Seminar on Industrial Control Systems: Analysis, Modeling and Computation (ITM Web of Conferences). 2016. vol. 6.

16. **Demidova L.A., Sokolova Ju.S.** Ispol'zovanie SVM-algoritma dlja utochnenija reshenija zadachi klassifikacii ob'ektov s primeneniem algoritmov klasterizacii (Use of the SVM algorithm for the decision specification of the classification problem of objects with application of the clustering algorithms), Vestnik Rjazanskogo gosudarstvennogo radiotehnicheskogo universiteta, 2015. no. 1 (51), pp. 103-113 (in Russian).

17. **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.

18. **Demidova L., Sokolova Yu**. SVM-Classifier Development With Use Of Fuzzy Clustering Algorithms' Ensemble On The Base Of Clusters' Tags' Vectors' Similarity Matrixes, 16th International Symposium on Advanced Intelligent Systems 2015, pp. 889-906.

19**. Demidova L., Sokolova Yu**. Training Set Forming For SVM Algorithm With Use Of The Fuzzy Clustering Algorithms Ensemble On Base Of Cluster Tags Vectors Similarity Matrices, 2015 International Conference «Stability and Control Processes» in Memory of V.I. Zubov (SCP), 2015, pp. 619-622.

20. **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, vol. 2, 2006, pp. 2126-2136.

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

22. **Jun Sun, Choi-Hong Lai, Xiao-Jun Wu**. Particle Swarm Optimisation: Classical and Quantum Perspectives. CRC Press, 2011, 419 p.