УДК 004.855.5
DOI: 10.46548/21vek-2021-1054-0019

АНСАМБЛЕВЫЕ МЕТОДЫ В ЗАДАЧЕ МНОКОКЛАССОВОЙ SVM-КЛАССИФИКАЦИИ

Костров Борис Васильевич, доктор технических наук, профессор,
заведующий кафедрой электронных вычислительных машин
Баранчиков Алексей Иванович, доктор технических наук, доцент,
профессор кафедры электронных вычислительных машин
Клюева Ирина Алексеевна, программист кафедры вычислительной и прикладной математики
Рязанский государственный радиотехнический университет имени В.Ф. Уткина
(390005, г. Рязань, ул. Гагарина, 59/1, e-mails: kostrov.b.v@evm.rsreu.ru,alexib@inbox.ru,i.klyueva-job@yandex.ru)

Аннотация. При решении задач интеллектуального анализа данных на практике достаточно распространенными являются задачи многоклассовой классификации, когда используются более двух классов объектов в исходных данных. Для решения подобных задач применяются различные методы построения ансамблей классификаторов. Процесс ансамблирования основывается на различных способах компенсации ошибок классификаторов, входящих в ансамбль. Ансамбли обладают рядом преимуществ, к которым относится высокая способность к обобщению. Однако существуют недостатки ансамблей алгоритмов, такие как: подверженность к переобучению, сложность выбора базовых алгоритмов, составляющих ансамбль, неэффективное использованию обучающей выборки. В настоящей работе рассматриваются методы ансамблирования с применением принципов стекинга и использования метаклассификации на основе реализации *SVM*-алгоритма. В представленных методах предлагаются различные подходы к генерации метахарактеристик при нескольких разбиениях обучающей выборки. При этом для генерации метахарактеристик используется весь объем обучающей выборки, а значения метахарактеристик определяются на основе нечеткой оценки принадлежности объектов к классам. Результаты экспериментальных исследований показывают, что разработанные методы обеспечивают лучшее качество классификации по сравнению с многоклассовой *SVM*-классификацией с использованием декомпозиционных подходов «один против всех» и «один против одного», многоклассовой классификацией с использованием классического метода стекинга и метода блендинга.

Ключевые слова: интеллектуальный анализ данных, машинное обучение, многоклассовая классификация, машина опорных векторов, ансамбли классификаторов, стекинг.

THE ENSEMBLE METHODS IN THE MULTI-CLASS SVM CLASSIFICATION PROBLEM

Kostrov Boris Vasilievich, doctor of technical sciences, professor, head of the Electronic ComputersDepartment
Baranchikov Aleksey Ivanovich, doctor of technical sciences, associate professor,
professor of the Electronic Computers Department
Klyueva Irina Alekseevna, programmer of the Computational and Applied MathematicsDepartment
Ryazan State Radio Engineering University named after V.F. Utkin
(390005, Ryazan, Gagarina street, 59/1, e-mails: kostrov.b.v@evm.rsreu.ru, alexib@inbox.ru,i.klyueva-job@yandex.ru)

Abstract. When solving data mining problems, in practice, multi-class classification problems are quite common, when more than two classes of objects are used in the source data. To solve such problems, various methods of constructing classifier ensembles are used. The process of ensembling is based on various ways to compensate for the errors of the classifiers included in the ensemble. Ensembles have a number of advantages, which include a high ability to generalize. However, there are disadvantages of ensembles of algorithms, such as: the susceptibility to overfitting, the complexity of choosing the basic algorithms that make up the ensemble, and the inefficient use of the training sample. In this paper, we consider the methods of ensembling using the principles of stacking and the use of metaclassication based on the implementation of the SVM algorithm. In the presented methods, various approaches to generating metacharacteristics with multiple partitions of the training sample are proposed. At the same time, when generating metacharacteristics, the entire volume of the training sample is used, and the values of metacharacteristics are determined based on a fuzzy assessment of the objects ' belonging to classes. The results of experimental studies show that the developed methods provide better classification quality compared to the multiclass SVM classification using the decomposition methods "one against all" and "one against one", the multiclass classification using the classical stacking method and the blending method.

Keywords: data mining, machine learning, multiclass classification, support vector machine, classifier ensembles, stacking.

Введение. В настоящее время технологии интеллектуального анализа данных (*data mining*) [1-4] востребованы и широко применяются в различных сферах деятельности, где накоплены ретроспективные данные. Технологии *data mining* ориентированы на поиск взаимосвязей, скрытых закономерностей в различных по своему объему и структуре массивах информации.

Для решения задач интеллектуального анализа данных применяются алгоритмы машинного обучения (*machine learning*) [5-7]. Задачей машинного обучения является предсказание результата по входным данным, чем разнообразнее входные данные, тем проще алгоритму машинного обучения определить закономерности в данных и обеспечить наиболее точный результат анализа. Востребованность алгоритмов машинного обучения объясняется их способностью к обучению на основе новых данных и поиску лучшего среди всех возможных решений при анализе и обработке большого объема информации.

В настоящее время машинное обучение имеет многочисленные сферы применения, включающие распознавание речи, обработку естественных языков, компьютерное зрение, диагностику заболеваний в медицине, биоинформатику, техническую диагностику производственного оборудования, скоринговые и экспертные системы и др.

Различают дедуктивное и индуктивное обучение. Если дедуктивное обучение относится к области экспертных систем, в которых содержатся формализованные знания, то основная задача индуктивного обучения заключается в восстановлении некоторой зависимости по эмпирическим данным. Различают три основных типа индуктивного обучения: обучение с учителем (*supervised learning*), обучение без учителя (*unsupervised learning*), обучение с подкреплением (*reinforcement learning*) и др.

Одной из распространенных задач машинного обучения в области интеллектуального анализа данных является задача классификации, решение которой основано на принципах обучения по прецедентам и предполагающая использование данных с заранее предопределенным конечным множеством классов. В базовом представлении задача классификации относится к бинарному типу, когда классифицируемые данные содержат два класса объектов. Однако на практике при классификации встречаются большие массивы данных, содержащие более двух классов объектов. Поэтому наиболее распространенным и при этом более сложным типом задач классификации является задача многоклассовой классификации.

Одним из известных алгоритмов обучения с учителем является *SVM*-алгоритм. Алгоритм машины опорных векторов (*Support Vector Machine, SVM*) [7-12, 15-21] достаточно популярен и востребован в области машинного обучения, поскольку позволяет оптимизировать построение разделяющей гиперплоскости (*OSH*) между двумя классами и обеспечивает разделение данных вне зависимости от их распределения в исходном пространстве характеристик (классификация в случаях отсутствия линейной разделимости объектов), при этом характеризуется высоким качеством обобщения [17]. *SVM*-алгоритм находит применение в различных областях интеллектуального анализа данных: анализ текста, распознавание фонем, классификация изображений, биоинформатика и т. д. [17].

Математическое описание *SVM*-алгоритма было первоначально разработано для бинарной классификации, ввиду чего актуальна проблема исследования для случая многоклассовой *SVM*-классификации.

SVM-классификатор может применяться при решении задачи многоклассовой классификации с использованием декомпозиционных подходов, обеспечивающих переход от задачи классификации на множество классов к нескольким задачам бинарной классификации: «один против всех» (*One-Against-All* (*OAA*), или *One-vs.-All* (*OvA*), или *One-vs.-Rest* (*OvR*), далее – *OvR*); «один против одного» (*One-vs.-One*, (*OvO*)) [8, 9]; *Pair wise Fuzzy* (*SVM*) [10]; *Directed Acyclic Graph Support Vector Machines* (*DAG*) [11]; *Weighted DAG SVM* (*WDAG SVM*) [12] и др.

Однако применение *SVM*-классификатора в сочетании с декомпозиционными методами зачастую не обеспечивает приемлемого качества многоклассовой классификации, в частности, если объекты разных классов сильно перемешены в наборе данных [12].

В машинном обучении благодаря гибкости и обобщающей способности широкое применение находят ансамбли алгоритмов. К популярным методам ансамблирования относятся такие методы как: бэггинг (*bagging*), бустинг (*boosting*) и стекинг (*stacked generalization*, *stacking*) [13, 14].

В задаче классификации под ансамблированием подразумевается использование композиции нескольких классификаторов с целью получения лучшего качества классификационного решения по сравнению с качеством решения каждого отдельно взятого классификатора. При этом предполагается, что за счет различных способов комбинирования и объединения ответов базовых классификаторов компенсируются ошибки каждого из них и достигается приемлемое качество решения в ансамбле.

Цель работы состоит в разработке методов повышения качества многоклассовой *SVM*-классификации на основе применения принципов ансамблирования.

С этой целью представлены четыре ансамблевых метода классификации, основанных на использовании идей стекинга и генерации метахарактеристик при нескольких разбиениях обучающей выборки, проведен сравнительный анализ разработанных методов с использованием различных методов ансамблирования алгоритмов: бэггинга, бустинга, стекинга, а также метода блендинга (*blending*) [13, 14].

Материалы и результаты исследования. Основная идея стекинга заключается в использовании разнородных базовых алгоритмов для применения их ответов в качестве метахарактеристик при обучении некоторого «обобщающего» алгоритма (метаалгоритма).

В разработанных методах в качестве метакласcификатора предлагается использовать композицию бинарных *SVM*-классификаторов на основе того или иного декомпозиционного метода решения задачи многоклассовой классификации: «один против всех», «один против одного».

В разработанных методах производится несколько

разбиений ($t=\overline{1,T}$) обучающей выборки на *K* блоков, после чего объединяются ответы базовых классификаторов для каждого *k*-го блока ($k=\overline{1,K}$). Таким образом, множество разбиений обучающей выборки представлено в виде $\{\{(Z_{tk}, Y_{tk}),\ k=\overline{1,K}\},\ t=\overline{1,T}\}$.

Пусть $b_{wt} \in B$ – базовый классификатор из множества $B=\{b_{11}, b_{12}, ..., b_{21}, b_{22}, ..., b_{wt}\}$, обеспечивающий генерацию метахарактеристик, построен на основе *w*-го ($w=\overline{1,W}$) базового алгоритма со значениями параметров, соответствующими *t*-му ($t=\overline{1,T}$) разбиению обучающей выборки; $B_t \subset B$ – подмножество классификаторов, используемое на *t*-м разбиении обучающей выборки.

1. DGM-1-метод. Для каждого *t*-го ($t=\overline{1,T}$) разбиения обучающей выборки реализуется построение одного базового классификатора из множества $B(b_{wt} \in B)$ и использование его ответов в качестве метахарактеристик для обучения другого базового классификатора при (*t+1*)-м разбиении обучающей выборки. Метаклассификатор обучается на основе исходного набора характеристик обучающей выборки и метахарактеристик, полученных при реализации последнего (*t=T*)-го разбиения обучающей выборки.

2. DGM-2-метод. Для каждого *t*-го ($t=\overline{1,T}$) разбиения обучающей выборки строится подмножество классификаторов $B_t \subset B$ и используется усреднение их ответов в качестве метахарактеристик для обучения другого множества базовых классификаторов $B_{t+1} \subset B$ при (*t+1*)-м разбиении обучающей выборки. Метаклассификатор обучается на основе исходного набора характеристик обучающей выборки и метахарактеристик, полученных при усреднении ответов базовых классификаторов на последнем (*t=T*)-м разбиении обучающей выборки.

3. IGM-1-метод. Для каждого *t*-го ($t=\overline{1,T}$) разбиения обучающей выборки с помощью кросс-валидации проводится оценка классификаторов из подмножества $B_t \subset B$ и выбор для генерации метахарактеристик классификатора с лучшим значением оценки перекрестной проверки. Метаклассификатор обучается на основе исходного набора характеристик обучающей выборки и метахарактеристик, полученных при усреднении ответов базовых классификаторов при различных *t*-ых разбиениях обучающей выборки.

4. IGM-2-метод. Для каждого *t*-го ($t=\overline{1,T}$) разбиения обучающей выборки с помощью кросс-валидации проводится оценка классификаторов из подмножества $B_t \subset B$. При (*t+1*)-м разбиении обучающей выборки увеличивается число построений классификаторов на основе того алгоритма, который использовался при построении классификатора с лучшей оценкой перекрестной проверки на предыдущем *t*-м разбиении обучающей выборки. Метаклассификатор обучается на основе исходного набора характеристик обучающей выборки и метахарактеристик, полученных при усреднении ответов базовых классификаторов при различных *t*-ых разбиениях обучающей выборки.

Таким образом, *DGM*-1 и *DGM*-2-методы реализуют обучение базовых классификаторов с использованием метахарактеристик на основе ответов базовых классификаторов с предыдущего разбиения обучающей выборки (*dependent generation of metacharacters, DGM*). При реализации *IGM*-1 и *IGM*-2-методов предполагается, что создание метахарактеристик происходит независимо на разных разбиениях обучающей выборки (*independent generation of metacharacters, IGM*). Оценка перекрестной проверки, используемая в *IGM*-1 и *IGM*-2-методах, представляет собой усредненную ошибку по всем *k* ($k=\overline{1,K}$) блокам на текущем *t*-м ($t=\overline{1,T}$) разбиении обучающей выборки.

При реализации разработанных методов многоклассовой классификации в качестве базовых алгоритмов выбраны различные ансамблевые алгоритмы на основе бэггинга и бустинга: алгоритм случайного леса (*random forest*), *AdaBoost*-, *XGBoost*-алгоритмы.

Сравнение разработанных методов проводилось на основе значений показателя сбалансированной точности (*balanced accuracy, BA*) для четырех тестовых наборов данных: 3 набора данных из репозитория *UCI Machine Learning Repository* и один набор данных соревнования по машинному обучению с платформы *Kaggle*.

Базовая формула расчета показателя для бинарной классификации имеет вид

$$BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \quad (1)$$

где *TP* – число истинно положительных наблюдений; *TN* – число истинно отрицательных наблюдений; *FP* – число ложноположительных наблюдений; *FN* – число ложноотрицательных наблюдений. Показатель сбалансированной точности *BA* отличается от показателя общей точности (общей доли правильных ответов) тем, что при его расчете учитывается доля объектов в каждом из классов. Соответственно показатель *BA* отражает наиболее достоверную оценку качества классификации наборов данных с несбалансированными классами [15].

На рисунке 1 представлены графики для тестовых наборов данных с отражением числа экспериментов, в которых соответствующий метод продемонстрировал лучшие результаты среди остальных разработанных методов. На рисунке 2 представлены результаты сравнительного анализа методов для всех проведенных экспериментов. Поскольку для каждого набора данных было произведено по 5 экспериментов, в идеале для каждого разработанного метода а) на рисунке 1 максимальное количество экспериментов по одному набору данных составило бы число 15, б) на рисунке 2 общее количество экспериментов составило бы в сумме число 60.

В таблице 1 приведены значения показателя сбалансированной точности при реализации многоклассовой классификации на основе различных ансамблевых методов. Результаты таблицы 1 позволяют утверждать, что по сравнению с аналогами (многоклассовой *SVM*-классификацией на основе декомпозиционных методов *OvR* и *OvO* (*SVM_ovo*, *SVM_ovr*),

многоклассовой классификацией с использованием методов бэггинга и бустинга (*RF, AdaBoost, XGBoost*), с многоклассовой классификацией на основе методов стекинга и блендинга (*stacking, blending*), предложенные методы многоклассовой классификации (*DGM*-1, *DGM*-2, *IGM*-1, *IGM*-2) обеспечивают лучшие значения показателя BA.

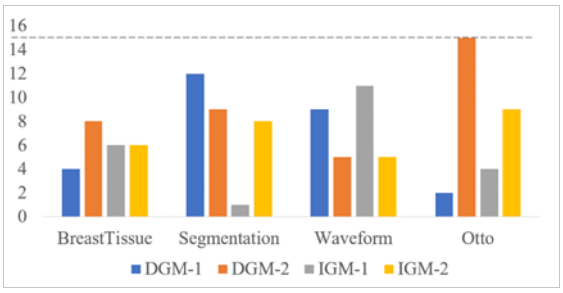

Рисунок 1 – Сравнительный анализ разработанных методов для тестовых наборов данных

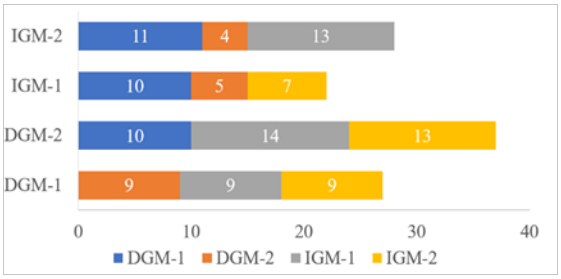

Рисунок 2 – Общий сравнительный анализ методов

Таблица 1 – Результаты многоклассовой классификации

Набор данных	SVM_ovo	SVM_ovr	RF	AdaBoost	XGBoost	blending	stacking	DGM-1	DGM-2	IGM-1	IGM-2
BreastTissue	83,05	83,05	90,64	89,66	-	84,83	92,33	95,19	97,56	95,89	96,25
Segmentation	92,92	92,92	93,26	88,59	-	95,84	95,50	97,30	96,90	96,23	97,07
Waveform	85,13	85,13	85,53	-	85,66	84,89	85,47	86,05	85,90	86,14	85,96
Otto	60,03	60,03	63,54	-	66,08	59,87	64,27	68,18	69,87	68,74	69,36

Заключение. К преимуществам разработанных методов многоклассовой классификации по сравнению с аналогами следует отнести использование нечеткой оценки принадлежности объектов к классам при получении значений метахарактеристик, проведение оценки базовых классификаторов посредством кросс-валидации.

По сравнению с методом стекинга в разработанных методах при получении значений метахарактеристик для объектов тестовой выборки проводится обучение базовых классификаторов на том же объеме обучающей выборки, что и при генерации метахарактеристик для объектов обучающего набора. В отличие от метода блендинга разработанные методы обеспечивают эффективное использование всего объема обучающей выборки при построении метаклассификатора.

СПИСОК ЛИТЕРАТУРЫ:

1. Дюк, В. А. Data mining: учебный курс / В. А. Дюк, А. П. Самойленко. – СПб.: Питер, 2001. – 368 с.
2. Барсегян, А. А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. Учеб. пособие. 2– е изд. / А. А. Барсегян, М. С. Куприянов, В. В. Степаненко, И. И. Холод. – СПб.: БВХ–Петербург, 2007. –384 с.
3. Дьяков, О. А. Особенности применения методов Data Mining в скоринговых решениях для коммерческих банков / О. А. Дьяков // Научные записки молодых исследователей. – 2017. – №3 – С. 5–11.
4. Загоруйко, Н. Г. Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. – Новосибирск: ИМ СО РАН, 1999. – 270 с.
5. Воронцов, К. В. Комбинаторные оценки качества обучения по прецедентам / К. В. Воронцов // Докл. РАН. – 2004. – Т. 394, № 2. – С. 175–178.
6. Воронцов, К. В. Комбинаторный подход к оценке качества обучаемых алгоритмов / К. В. Воронцов // Математические вопросы кибернетики / Под ред. О. Б. Лупанов. – М.: Физматлит. – 2004. – Т. 13. – С. 5–36.
7. Vapnik, V. Statistical Learning Theory / V. Vapnik. – New York: John Wiley & Sons, 1998. – 734 p.
8. Hsu, C. W. A comparison of methods for multiclass support vector machines / C. W. Hsu, C. J. Lin // IEEE Trans. NeuralNetworks. – 2002. – 13 (2) – P. 415-425.
9. Huang, P. X. Individual feature selection in each One-versus-One classifier improves multi-class SVM performance / P. X. Huang, R. B. Fisher // Reference Source. – 2014. – P. 98-103.
10. Abe, S. Fuzzy Support Vector Machines for Multiclass Problems / S. Abe, T. Inoue // ESANN'2002 proceedings – European Symposium on Artificial Neural Networks. – Bruges (Belgium), 2002. – P. 113–118.
11. Platt, J. Large margin DAGs for multiclass classification / J. Platt, N. Cristianini, J. Shawe-Taylor // Advances in Neural Information Processing Systems. – 2000. – Vol. 12. – P. 547–553.
12. Sabzekar, M. Improved DAG SVM: A New Method for Multi-Class SVM Classification / M. Sabzekar, M. Ghasemigol, M. Naghibzadeh, H. S. Yazdi // Int'l Conf. ArtificialIntelligence (ICAI'09). – Las Vegas (USA), 2009. – Vol. 2. – P. 548–553.
13. Гущин, А. Е. Методы ансамблирования обучающихся алгоритмов [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/images/5/56/Guschin2015Stacking.pdf(датаобращения 03.06.2021).
14. Дьяконов, А. Г. Стекинг (Stacking) и блендинг (Blending) [Электронный ресурс]. – Режим доступа: https://dyakonov.org/2017/03/ (дата обращения 03.06.2021).
15. Klyueva, I. The two-stage classification based on 1-SVM and RF classifiers / L. Demidova, I. Klyueva // Journal of Physics: Conference Series. – 2020. – Vol. 1727. – P. 012007.
16. Демидова, Л. А. Разработка ансамблей SVM-классификаторов / Л. А. Демидова, Ю. С. Соколова // Математические методы и информационные технологии в экономике, социологии и образовании: сборник статей XXXIV Международной научно-технической конференции. – Пенза: Приволжский Дом знаний, 2015. – С. 57–61.
17. Клюева, И. А. Повышение качества многоклассовой SVM-классификации на основе feature engineering / И. А. Клюева // Cloud of science. – 2020. – Т. 7, № 1. – С. 207–218.
18. Klyueva, I. Improving Quality of the Multiclass SVM Classification Based on the Feature Engineering / I. Klyueva // 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA–2019). – Lipetsk, 2019. – P. 491–494.
19. Klyueva, I. Intellectual Approaches to Improvement Of the Classification Decisions Quality On the Base Of the SVM Classifier / L. Demidova, I. Klyueva, Y. Sokolova, N. Stepanov, N. Tyart // Procedia Computer Science. – 2017. – Vol. 103. – P. 222–230.
20. Klyueva, I. Hybrid approach to improving the results of the SVM classification using the random forest algorithm / L. Demidova, I. Klyueva, A. Pylkin // Procedia Computer Science. – 2019. – Vol. 150. – P. 455–461.
21. Клюева, И. А. Алгоритм случайного леса в задаче повышения качества SVM-классификации / Л. А. Демидова, И. А. Клюева // Вестник Рязанского государственного радиотехнического университета. – 2018. – № 65. – С. 74–83.

Статья поступила в редакцию 16.05.2021
Статья принята к публикации 16.06.2021