Двух-шаговый метод построения неоднородного ансамбля классификаторов для прогнозирования банкротства

Ю.А. Зеленков
д.т.н., Финансовый университет при Правительстве РФ, yuri.zelenkov@gmail.com

Е.А. Федорова
д.э.н., Высшая школа экономики, ecolena@mail.ru

Аннотация

Предлагается двух-шаговый метод классификации, основанный на генетических алгоритмах. На первом шаге обучаются классификаторы различных моделей, на втором они объединяются в ансамбль. Отбор значимых переменных для каждого классификатора на первом шаге и определение весов в ансамбле на втором шаге производится при помощи генетических алгоритмов. Характеристики предложенного метода проверены на сбалансированной выборке данных о российских компаниях, включающей 912 наблюдений (456 - банкроты и 456 – успешные компании) и 55 признаков (финансовые коэффициенты и факторы внешней среды предприятий на макро и микроуровнях), которая была разделена на обучающее и тестовое подмножество в пропорции 80/20. На тестовой выборке предложенный метод показал наилучшее значение точности предсказания (accuracy = 0,934) среди всех проверенных методов построения ансамблей. Он также продемонстрировал сбалансированное соотношение показателей precision и recall, т.е. он с достаточно высокой точностью обнаруживает банкротов (recall = 0.953) и в то же время достаточно редко ошибается, классифицируя успешные компании как банкроты (precision = 0.910). Также проверена способность метода к выделению признаков, значимых с точки зрения решаемой задачи. При исключении признаков, которые оказались значимыми при построении менее чем 50% классификаторов, участвующих в ансамбле, все характеристики метода улучшаются (accuracy = 0.951, precision = 0.932, recall = 0.965).

Ключевые слова: банкротство предприятий; модели прогнозирования банкротства; ансамбли классификаторов; отбор признаков; генетический алгоритм.

1. Введение и обзор литературы

Проблема прогнозирования банкротства занимает особое место среди практических и теоретических вопросов управления компанией. Все исследования данной сферы условно могут быть поделены на две группы. **К первому направлению** относятся работы, касающиеся выбора *набора независимых признаков*, обеспечивающего высокую точность

прогнозирования (подробный анализ см. Ravi Kumar and Ravi, 2007). При этом чаще всего исследуются показатели финансового состояния предприятий (Fedorova et al., 2013), также рассматриваются факторы корпоративного управления (Liang et al., 2016), внешней среды (Tinoco & Wilson, 2013), уровень развития законодательства (Rowoldt & Starke, 2016) и т.д.

Вторая группа исследований фокусируется на *методологии прогнозирования.* Обычно задачу прогнозирования банкротств рассматривают как проблему классификации: имеется множество объектов одного типа, для которых известно к каким классам они относятся, на его основе требуется построить алгоритм, способный классифицировать аналогичные объекты, класс принадлежности которых неизвестен. Для решения этой задачи используются как традиционные методы классификации: логистическая регрессия (LR), k ближайших соседей (kNN), метод опорных векторов (SVM), наивный баейсовский классификатор (NB), деревья классификации (DT), многослойные персептроны (MLP), так и различные подходы, основанные на построении ансамблей классификаторов. Сравнительному анализу эффективности различных методов классификации при решении задачи прогнозирования банкротств посвящены многочисленные работы (Liang et al., 2016; Fedorova et al., 2013; Tsai et al., 2014; Peng et al., 2011).

Помимо в той или иной степени детерминированных методов построения классификаторов некоторые исследователи используют *эволюционные техники* для оптимизации их параметров. При построении ансамблей, эволюционные техники позволяют получить системы, более точно учитывающие особенности предметной области, поскольку они обеспечивают большую диверсификацию классификаторов (Brown et al. 2005; Kim &Kang, 2012).

Большинство методов построения ансамблей оперируют классификаторами одного типа. Однако гетерогенные ансамбли превосходят традиционные подходы, опирающиеся на однородные классификаторы, поскольку используют для принятия решений комбинацию различных правил и способов извлечения информации (Wozniak et al., 2014). Подход, основанный на построении ансамбля из классификаторов различного типа получил название Multi-Classifier Systems (MCS) и активно развивается.

Общей проблемой при решении задачи классификации является *отбор признаков*, релевантных исследуемой проблеме (Han et al, 2012; Mukhopadhyay et al, 2014). Liang et al, (2015) отметили, что общепризнанное соглашение о факторах, которые могут служить входами моделей для предсказания финансовых проблем, отсутствует, поэтому многие исследователи рассматривают отбор признаков как обязательную часть подготовки данных. В настоящее время разработано большое число техник отбора признаков, значительную часть из них занимают методы, основанные на имитации природы:

генетические алгоритмы (GA), метод роя частиц, симуляция отжига и т.д. (Guyon & Elisseeff, 2003; Liang et al. 2015). Эволюционные техники позволяют с наименьшими затратами исследовать пространство параметров и выбрать оптимальный набор параметров для каждого индивидуального классификатора.

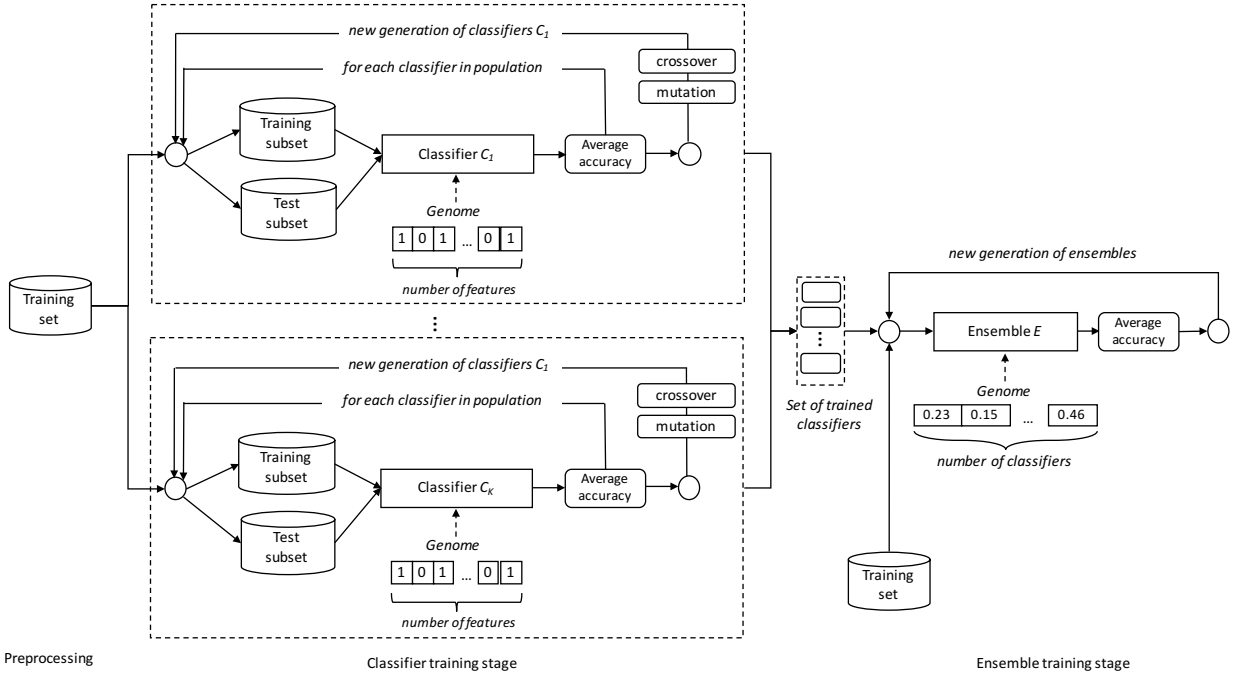

Рис. 1. Алгоритм построения гибридного ансамбля

2. Метод классификации

Алгоритм конструирования гибридного ансамбля должен поддерживать включение взаимно комплементарных базовых классификаторов, которые обеспечивают высокое разнообразие и, следовательно, точность. На пути к созданию такого алгоритма возникают две проблемы. Во-первых, отсутствуют общепринятые методы измерения разнообразия (Brown et al. 2005). Во-вторых, невозможно заранее предсказать какие классификаторы могут быть комплементарны, это зависит от типа решаемой задачи и конкретного набора данных.

Еще одна проблема связана с выбором архитектуры ансамбля классификаторов. Мы используем голосующий ансамбль (voting classifier, VC), поскольку это наиболее простой способ реализовать MCS. Результирующее значение ансамбля формируется как взвешенная сумма ответов индивидуальных классификаторов. Класс принадлежности каждого объекта определяется значениями 1 (банкрот) и -1 (не банкрот). Знак выхода ансамбля соответствует классу принадлежности объекта, модуль выхода – степени уверенности.

Предлагаемый алгоритм реализуется в 2 стадии (рис. 1). На первой стадии производится обучение индивидуальных классификаторов и отбор наиболее адекватного

множества признаков для каждого из них. Поскольку, как отмечено выше, заранее невозможно предсказать, какие классификаторы будут обладать взаимодополняющими свойствами, на этой фазе исследователь может использовать любые доступные модели.

2.1. Первая стадия. Обучение набора классификаторов.

Классификатор кодируется массивом *G* длиной *N*, который описывает признаки, используемый для его обучения (*N* – количество признаков в исследуемом наборе данных). Элементы массива могут принимать только значения 0 или 1. Если элемент равен 0, соответствующий признак исключается из обучающей выборки. Начальная популяция создается из классификаторов одного типа, причем всем элементам генотипа *G* каждой особи присваивается значение 1. Тем самым обучение каждого классификатора начинается с полного набора признаков.

Лучшая особь всегда копируется в новую популяцию без изменений (принцип элитизма). Отбор остальных особей производится на основании их ранга. Операция мутации применяется к случайно выбранному элементу генотипа *G* отобранной особи, при этом его значение заменяется на противоположное, т.е. 0 заменяется на 1, а 1 на 0. При кроссовере производится обмен случайно выбранной подстрокой между двумя отобранными особями.

Приспособленность каждой особи вычисляется как среднее значение accuracy классификации:

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Здесь TP - количество верно классифицированных объектов банкротов, FP – количество небанкротов, отнесенных к банкротам, TN - количество верно классифицированных небанкротов, FN - количество банкротов, отнесенных небанкротам.

На первой стадии обучения исследователь задает множество *M* типов классификаторов, которые будут использованы для построения ансамбля. Классификатор каждого типа обучается по описанному выше алгоритму, набор обученных классификаторов передается на следующую стадию.

2.2. Вторая стадия. Построение ансамбля

На второй стадии из набора обученных на первой стадии классификаторов конструируется голосующий ансамбль. Ансамбль кодируется массивом *w* вещественных чисел больших, либо равных нулю размером *M*, которые задают значение весового коэффициента при соответствующем классификаторе. При создании начальной популяции

элементы w_i, $i = 1 \ldots M$ задаются случайным образом, при этом соблюдается условие нормальности $\sum_i w_i = 1$.

Правила отбора те же, что и на первой стадии: элитизм, отбор по рангу, аналогичные условия применения мутации и кроссовера. Операция мутации применяется ко всем элементам массива *w* отобранной особи, их значение изменяется на случайную равномерно распределенную величину из диапазона (-0.1; 0.1). Если в результате получено отрицательное значение w_i, оно заменяется на 0. При этом условие нормальности не соблюдается. Данные параметры были определены во время экспериментальных запусков алгоритма.

Приспособленность ансамбля вычисляется как accuracy на обучающей выборке. Класс объекта C_E вычисляется как взвешенная сумма выходов отдельных классификаторов $C_E = \sum_{i \in M} w_i p_i$, где p_i – выход i-го классификатора. Если знак C_E совпадает с заданным типом объекта, то объект считается распознанным правильно. Модуль значения C_E свидетельствует о степени уверенности при классификации.

Программный код, реализующий метод, был разработан на языке python на основе библиотеки машинного обучения scikit-learn (Pedregosa et al. 2011).

3.Анализ данных

Для оценки качества предложенного метода был использована сбалансированная выборка данных о российских компаниях, включающая 912 наблюдений (из них 456 - банкроты и 456 – успешные компании) и 55 признаков. Данная выборка была разбита на 2 множества – обучающее и тестовое в пропорции 80/20, которые включали, соответственно, 729 и 183 наблюдения.

Список признаков включал стандартные финансовые коэффициенты (ликвидности, финансовой устойчивости, оборачиваемости и рентабельности и т.д.), показатели модели Альтмана (Altman, 1968), а также финансовые коэффициенты, наиболее часто встречающиеся в исследованиях банкротств в различных странах (Giacomino, Bellovary & Akers, 2007). В набор признаков также были включены внешние факторы, характеризующие экономическую конъюнктуру и деловой климат (Федорова и др., 2016).

4. Результаты

В таблице 1 представлены результаты обучения базовых классификаторов (первая стадия метода), при этом для всех классификаторов использовались следующие параметры: размер популяции – 50, число поколений – 50. Точность до обучения рассчитана до запуска генетического алгоритма (при этом набор признаков *N* = 55, т.е. включает все элементы). В

качестве базовых классификаторов использованы широко известные модели kNN, LR, NB, DT и SVM с параметрами, которые в библиотеке scikit-learn используются по умолчанию.

Таблица 1. Результаты обучения базовых классификаторов

Модель	Обучающая выборка					Тестовая выборка			Весовой коэффициент в ансамбле w_i
	До обучения ($N = 55$)	После обучения							
	Accuracy	Accuracy	Precision	Recall	N	Accuracy	Precision	Recall	
kNN	0.830	0.871	0.851	0.911	34	0.831	0.787	0.871	0.283
LR	0.808	0.829	0.810	0.868	51	0.825	0.773	0.882	0.013
NB	0.557	0.602	0.598	0.669	33	0.579	0.540	0.635	0.122
DT	0.878	0.892	0.895	0.892	46	0.847	0.852	0.812	0.184
SVM	0.848	0.855	0.839	0.884	54	0.845	0.816	0.856	0.287

Отметим, что при обучении классификаторы выбирают разное количество значимых признаков N, метод SVM выбирает самое большое количество, NB - наименьшее количество.

Значения метрик precision and recall вычисляются по следующим формулам:

$$precision = \frac{TP}{TP + FP}; \quad recall = \frac{TP}{TP + FN}$$

В данном случае показатель precision характеризует ошибки классификатора, связанные с отнесением успешных компаний к банкротам, recall - ошибки, связанные с отнесением банкротов к успешным компаниям.

Также в таблице 1 приведены значения accuracy, precision and recall классификации на тестовой выборке. И на обучающей, и на тестовой выборке самую высокую точность демонстрирует метод DT.

Второй шаг - обучение ансамбля. При этом использовались следующие параметры: размер популяции – 40, число поколений – 40. Коэффициенты w_i полученного ансамбля приведены в крайней правой колонке таблицы 1.

В библиотеке scikit-learn реализовано значительное количество методов построения ансамблей. В таблице 2 представлены результаты сравнения этих методов с предлагаемым здесь на описанных выше обучающей и тестовой выборках. Среди методов, результаты которых представлены, пять bagging методов (на основе классификаторов DT, kNN, SVM, NB и LR), три метода AdaBoost (на основе классификаторов DT, SVM и NB), gradient boosting, методы стохастического построения ансамблей классифицирующих деревьев – RandomForest и ExtraTrees и метод построения voting classifier.

Таблица 2. Точность прогнозирования различных методов построения ансамблей

Метод	Training set			Test set		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
BAGGING - DT	0.999	0.997	0.999	0.913	0.906	0.906
BAGGING - KNN	0.962	0.939	0.989	0.923	0.874	0.976
BAGGING - SVM	0.909	0.902	0.922	0.918	0.880	0.953
BAGGING - NB	0.658	0.657	0.687	0.497	0.467	0.576
BAGGING - LR	0.896	0.889	0.908	0.880	0.846	0.906
ADABOOST - DT	0.999	1.000	0.999	0.880	0.846	0.906
ADABOOST - SVM	0.509	0.509	0.999	0.464	0.464	0.999
ADABOOST - NB	0.824	0.846	0.801	0.776	0.762	0.753
Gradient Boosting	0.984	0.969	0.999	0.858	0.824	0.882
Voting Classifier	0.929	0.916	0.946	0.891	0.857	0.918
RandomForest	0.999	0.999	0.999	0.831	0.855	0.765
ExtraTrees	0.999	0.999	0.999	0.929	0.901	0.953
Предложенный метод	0.963	0.943	0.987	0.934	0.910	0.953

На обучающей выборке предложенный метод демонстрирует несколько худшие характеристики чем некоторые другие методы, но на тестовой – превосходит всех. Это означает, что он гораздо успешнее избегает переобучения за счет сочетания техник random sampling и feature selection на шаге обучения индивидуальных классификаторов. Кроме того, предложенный метод демонстрирует достаточно сбалансированное соотношение показателей precision и recall, т.е. он с достаточно высокой точностью обнаруживает банкротов (recall = 0.953) и в то же время достаточно редко ошибается, классифицируя успешные компании как банкроты (precision = 0.910).

Отметим что предложенный метод обладает еще одним важным преимуществом – он позволяет отобрать оптимальный набор признаков для решения поставленной проблемы.

В список самых значимых признаков, которые были отобраны всеми 5 использовавшимися базовыми классификаторами, попали 4 коэффициента финансового состояния предприятия (текущая ликвидность, финансовая независимость, оборотные активы, коэффициент автономии). «Классические» модели прогнозирования банкротств используют только финансовые показатели, однако для развивающихся стран невозможно решить эту задачу, основываясь только на внутренних финансовых коэффициентах. В нашем случае в список наиболее значимых включены 6 факторов внешней среды. Среди них 3 признака, характеризующих макроэкономическую ситуацию: темп отзыва лицензий кредитных организаций, индекс РТС, индекс цен промышленных товаров. Все эти факторы выступают индикаторами кризисных ситуаций, которые провоцируют дополнительный

стресс для предприятий. Значимыми оказались также и 3 микроэкономических индикатора: доля рынка, факт недобросовестности поставщиков и уровень среднемесячной заработной платы.

В следующий по значимости набор признаков (признаны значимыми 4 классификаторами из 5) попали 18 факторов финансового состояния предприятия и 13 факторов внешней среды предприятия. На макроуровне значимым оказались показатели фондовых индексов, цена на нефть, ВВП. Кроме того, значимыми оказались микроэкономические факторы: показатели уровня конкуренции (степень монополизации, является ли предприятие естественной монополией или находится под государственным контролем), наличия проблем с поставщиками, а также количество исков к предприятию. Это подтверждает результаты других исследователей (см., например, Tinoco & Wilson, 2013) - объединение индикаторов внешней среды и финансовых показателей повышает прогностическую способность.

Для оценки способности предложенного метода решать задачу отбора признаков были исследованы три варианта выборок, включающих, соответственно 51 (признаны значимыми, как минимум, 3 классификаторами из 5 базовых), 41 (признаны значимыми, как минимум, 4 классификаторами) и 10 (признаны значимыми всеми 5 классификаторами) признаков. Данные тестирования на этих выборках приведены в таблице 3. При исключении признаков, которые оказались значимыми при построении менее чем 50% классификаторов, участвующих в ансамбле, все характеристики метода возрастают. При дальнейшем сокращении числа признаков характеристики ухудшаются.

Таблица 3. Тестирование качества отбора признаков

Кол-во признаков	Обучающая выборка			Тестовая выборка		
	Accuracy	Precision	Recall	Accuracy	Precision	Recall
55	0.963	0.943	0.987	0.934	0.910	0.953
51	0.971	0.992	0.989	0.951	0.932	0.965
41	0.956	0.929	0.989	0.929	0.900	0.953
10	0.937	0.909	0.973	0.918	0.880	0.953

Заключение

Предложенный метод обладает двумя преимуществами, во-первых, он обеспечивает точность классификации выше, чем другие известные методы построения ансамблей, во-вторых, может использоваться как инструмент выделения значимых факторов. Однако, выбор правила отсечения незначимых признаков должен исследоваться для каждой задачи.

Литература

1. Altman, E.I. (1968) Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. Journal of Finance. 23: 589-609.
2. Brown, G., Wyatt, J., Harris, R., Yao., X. (2005) Diversity creation methods: A survey and categorisation. Information Fusion. 6 (1): 5-20.
3. Fedorova, E., Gilenko, E., Dovzhenko, S. (2013) Bankruptcy prediction for Russian companies: Application of combined classifiers. Expert Systems with Applications. 40: 7285-7293
4. Giacomino, D.E., Bellovary, J.L., Akers M.D. (2007) A review of bankruptcy prediction studies: 1930 to present // Journal of Financial Education.1-42.
5. Guyon, I., Elisseeff, A. (2003) An Introduction to Variable and Feature Selection. Journal of Machine Learning Research. 3: 1157-1182.
6. Han, J., Kamber, M., Pei, J. (2012) Data mining: Concepts and Techniques. Waltham. MA. USA: Morgan Kaufmann.
7. Hernandez Tinoco, M., Wilson, N. (2013) Financial distress and bankruptcy prediction among listed companies using accounting, market and macroeconomic variables. International Review of Financial Analysis. 30: 394–419
8. Kim, M.-J., Kang, D.-K. (2012) Classifiers selection in ensembles using genetic algorithms for bankruptcy prediction. Expert Systems with Applications. 39: 9308-9314
9. Liang, D., Lu. C.-C., Tsai, C.-F., Shih. G.-A. (2016) Financial ratios and corporate governance indicators in bankruptcy prediction: A comprehensive study. European Journal of Operational Research (article in press).
10. Liang, D., Tsai, C.-F., Wu, H.-T. (2015) The effect of feature selection on financial distress prediction. Knowledge-Based Systems. 73: 289-297.
11. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., Coello Coello, C.A. (2014) A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I. IEEE Transaction on Evolutionary Computation. 18(1): 4-19.
12. Pedregosa, F. et al. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research. 12: 2825-2830.
13. Peng, Y., Wang, G., Kou, G., Shi, Y. (2011) An empirical study of classification algorithm evaluation for financial risk prediction. Applied Soft Computing. 11 (2): 2906–2915
14. Ravi Kumar, P., and Ravi, V. (2007) Bankruptcy Prediction in Banks and Firms via Statistical and Intelligent Techniques – A Review. European Journal of Operational Research. 180: 1–28.
15. Rowoldt, M., Starke, D. (2016) The role of governments in hostile takeovers – Evidence from regulation. anti-takeover provisions and government interventions. International review of Law and Economics. 47: 1-15.
16. Tsai, C.-F., Hsu, Y.-F., Yen., D.C. (2014) A comparative study of classifier ensembles for bankruptcy prediction. Applied Soft Computing. 24: 977-984.
17. Wozniak, M., Grana, M., Corchado, E. (2014) A survey of multiple classifier systems as hybrid systems. Information Fusion. 16 (1): 3-17.
18. Федорова Е. А., Зеленков Ю. А., Чекризов Д. В., Добрянская П. С. (2016). Влияние корпоративной культуры на банкротство российских предприятий на основе метода Partial Least Squares Path Modeling. Корпоративные финансы. 2 (38): 108-123.