

Синергия алгоритмов классификации (SVM Multimodelling)

С. Иванычев, А. Адуенко

sergeyivanichev@gmail.com, aduenko1@gmail.com

Московский физико-технический институт

В данной статье рассматривается проблема агрегирования небольшого количества сильных классификаторов с целью улучшения решений задач классификации и регрессии. В качестве примера подобной системы рассматривается система SVM алгоритмов использующая kernel-trick с различными ядрами. Для комбинации решений и улучшения качества прогнозирования в задачах классификации и регрессии (SVR) авторы предлагают способ формирования новых признаков на основе сгенерированных отступов (*margins*) каждым классификатором, приводят алгоритм обучения на полученных объектах и анализируют отличия множеств опорных объектов для различных ядер. В качестве практической проверки были проведены эксперименты на различных реальных данных из репозитория UCI.

Ключевые слова: *двухклассовая классификация, композиция алгоритмов, SVM, SVR, бэггинг*

Введение

Работа посвящена комбинированию небольшого количества сильных SVM, использующих kernel-trick с различными ядрами и получению агрегированного классификатора для улучшения решений задач классификации и регрессии.

SVM(Support Vector Machine) или *метод опорных векторов* [1, 2] — это один из наиболее распространенных и эффективных методов в машинном обучении, которые используются для задач классификации и регрессии (SVR). Задача математического программирования сводится к двойственной задаче, функционалы в которой не зависят от векторов признаков как таковых, а лишь от их попарных скалярных произведений [3]. Использование особых функций, *ядер*, то есть скалярных произведений в сопряженном пространстве, позволяет получить разделяющие поверхности между классами более сложной формы [4]. Наша цель — скомбинировать SVM с различными примененными ядрами для улучшения решения, а также анализ множеств опорных объектов в случае разных использованных ядер.

Наиболее классическими методами агрегирования алгоритмов являются бэггинг (*bagging*) [5] и бустинг (*boosting*) [6], и их вариации, однако они работают только с большим количеством слабых классификаторов, что делает невозможным использование его использование для указанного множества базовых алгоритмов.

Среди способов агрегации для небольшого количества классификаторов можно выделить, например, выбор большинства классификаторов [7], комбинирование ранжирований (*rankings*) по классам, сделанных различными классификаторами [8]. В дальнейшем было показано, что все подобные методы есть особые случаи составного классификатора из [9], появляющиеся при особых условиях или способах аппроксимации.

Различные способы агрегации SVM используются во многих задачах анализа данных. [10] использовали совокупность SVM для уменьшения ошибочно негативных классификаций (FP) в задаче фильтрации спама среди электронных писем. Для этого на электронных письмах были введены различные метрики, для каждой из них был приспособлен SVM, а затем результат получался голосованием [9]. [11], решавшие задачу распознавания

написанных рукой символов, делили множество признаков на четыре непересекающихся подмножества, и на каждом из них обучали SVM, увеличив этим самым коэффициент распознавания по сравнению с одним SVM.

В последнее время стал набирать популярность *метод многоядерного обучения* (MKL, multiple kernel learning) [12, 13, 14], который основывается на том, что линейная комбинация ядер также является ядром. Данный метод хорош при объединении данных из нескольких источников и полной автоматизации, так как суперпозиция функций может быть оптимизирована любым методом валидации (например кросс-валидацией).

Мы также предлагаем использовать накопившийся банк ядер, однако не на этапе обучения SVM, а на этапе агрегирования обученных алгоритмов. Известно, что алгоритм b_i для объекта x_j обучающей выборки генерирует *отступ* (margin). По отступу в общем случае можно определить не только предсказанный класс, но и насколько «уверен» в своем решении алгоритм. В случае банка с n ядрами и обучающей выборки с m сэмплами мы получим матрицу отступов $M \in \mathbb{R}^{m \times n}$. Отнормировав ее, мы получим новую матрицу «объект-признак», где вектором признаков каждого объекта будет вектор отнормированных отступов.

В этой работе предложен алгоритм обучения на матрице отступов, проведен анализ опорных объектов, генерируемые различными ядрами, а также проведено тестирование полученного алгоритма на реальных данных репозитория UCI.

Постановка задачи

Пусть $X^l = (\mathbf{x}_i, y_i)_{i=1}^l$ — обучающая выборка, $\mathbf{x} \in \mathbb{R}^n, y \in Y$ где $|Y| < \infty$ — задача классификации, $|Y| = \infty$ — задача регрессии. В данной статье под s -й *моделью* будем понимать SVM с ядром K_s где выбрано множество ядер:

$$\mathcal{K} = \{K_i\}_{i=1}^m$$

При обучении каждая модель дает классификатор или регрессор (в зависимости от типа Y). Например, для случая $Y \in \{-1, +1\}$ классификации алгоритм выглядит следующим образом:

$$b_s(\mathbf{x}) = \text{sign} \sum_{i=1}^l \lambda_i y_i K_s(\mathbf{x}_i, \mathbf{x}) - w_0$$

Где $\{\lambda_i\}$ и w_0 находятся из решения задачи математического программирования[4]

$$\begin{cases} \sum_{i=1}^l \lambda_i + \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \lambda_i \lambda_j y_i y_j K_s(\mathbf{x}_i, \mathbf{x}_j) \rightarrow \min_{\lambda} \\ 0 \leq \lambda_s \leq c, \quad i = 1 \dots l \\ \sum_{i=1}^l \lambda_i y_i = 0 \end{cases}$$

Побочным результатом обучения является то, что модель для обучающей выборки генерирует вектор *отступов* (margins) для каждого объекта.

Если мы рассмотрим множество моделей, обученных на X^l , то мы получим матрицу отступов размерности $M \in \mathbb{R}^{l \times m}$, в котором (i, j) -й элемент — это отступ i -го объекта в SVM с j -м ядром.

Пусть M — матрица «объект-признак», \mathcal{A} — множество алгоритмов вида

$$\mathcal{A} = \{a(\mathbf{x}) = g(\mathbf{x}, \theta) | \theta \in \Theta\} \quad g : \mathbb{R}^m \rightarrow Y \quad (1)$$

Пару (g, \mathcal{K}) назовем *мультимоделью*. $\mathcal{L}(y, y^*)$ — функционал качества, тогда перед нами стоит задача минимизации

$$L(y, g(M, \theta)) \rightarrow \min_{\mathcal{A}, \Theta} \quad (2)$$

Старая постановка задачи

Пусть $\mathcal{K} = \{K_i\}_{i=1}^m$ — множество ядер, выбранных для мультимоделирования SVM, $X^l = (\mathbf{x}_i, y_i)_{i=1}^l$ — обучающая выборка, $\mathbf{x} \in \mathbb{R}^n, y \in Y$. Тогда множество обученных SVM на данной выборке:

$$\mathcal{B} = \left\{ b_i | b_i = b_i(\mathbf{x}) = \text{sign} \sum_{i=1}^l \lambda_i y_i K_i(\mathbf{x}_i, \mathbf{x}) - w_0 \right\}_{i=1}^m \quad (3)$$

Где λ_s находятся из решения задачи математического программирования

$$\begin{cases} \sum_{i=1}^l \lambda_i + \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \lambda_i \lambda_j y_i y_j K_s(\mathbf{x}_i, \mathbf{x}_j) \rightarrow \min_{\lambda} \\ 0 \leq \lambda_s \leq c, \quad i = 1 \dots l \\ \sum_{i=1}^l \lambda_i y_i = 0 \end{cases}$$

И $w_0 = \sum_{i=1}^l \lambda_i y_i K_s(\mathbf{x}_i, \mathbf{x}_j) - y_j$ для такого j , что $\lambda_j > 0, M_j = 1$. Паре «выборка-обученные алгоритмы» соответствует матрица отступов размерности $M \in \mathbb{R}^{l \times m}$, в котором (i, j) -й элемент — это отступ i -го объекта в SVM с j -м ядром. Пусть M — матрица «объект-признак», \mathcal{A} — множество алгоритмов вида

$$\mathcal{A} = \{a(\mathbf{x}) = g(\mathbf{x}, \theta) | \theta \in \Theta\} \quad g : \mathbb{R}^m \rightarrow Y \quad (4)$$

$\mathcal{L}(y, y^*)$ — функционал качества, тогда перед нами стоит задача минимизации

$$L(y, g(M, \theta)) \rightarrow \min_{\mathcal{A}, \Theta} \quad (5)$$

Базовый вычислительный эксперимент

В данной секции попробуем установить связь между корреляцией векторов отступов, генерируемых разными ядрами на одной и той же обучающей выборке и степенью «похожести» ядер. Понятие «похожести» или близости отождествляется с введением метрики на пространстве ядер. Эту метрику терминах конкретной обучающей выборки можно задать, например, нормализованной симметрической разностью множеств опорных объектов

$$\rho_{X^l}(K_i, K_j) = \frac{\# [SV_i \Delta SV_j]}{\# [SV_i \cup SV_j]}$$

Где под SV_j понимается множество опорных объектов на j -м ядре, под знаком Δ — симметрическая разность. В качестве исходных данных возьмем датасеты Wine [15], German credit data [16] и Heart disease [17]. Варьируя коэффициент L2-регуляризации, в качестве базового набора ядер возьмем

- Линейное
- Полиномиальное (степени 3, 4, 5)
- RBF-ядро ($\gamma \in \{0.0001, 0.001, 0.01, 0.1, 1\}$)

Для каждого значения коэффициента регуляризации построим графики опишем полученное множество классификаторов.

German credit dataset

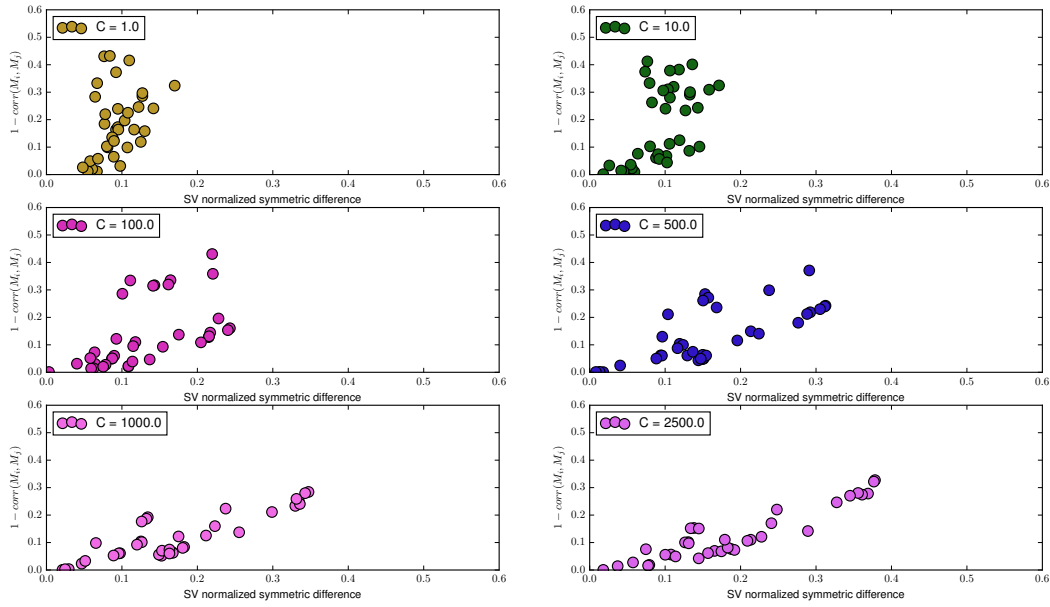


Рис. 1. German credit

Таблица 1. German info

	mean(#SV)	mean($1 - \text{corr}(M_i, M_j)$)	mean($\rho_{X^I}(K_i, K_j)$)	Correlation
$C = 1.0$	603.4	0.184	0.094	0.376
$C = 10.0$	603.6	0.187	0.097	0.537
$C = 100.0$	594.7	0.134	0.131	0.556
$C = 500.0$	584.3	0.133	0.161	0.717
$C = 1000.0$	581.6	0.120	0.172	0.870
$C = 2500.0$	577.9	0.126	0.189	0.918

Обучающая выборка состоит из 1000 семплов по 24 числовых признака.

Wine dataset

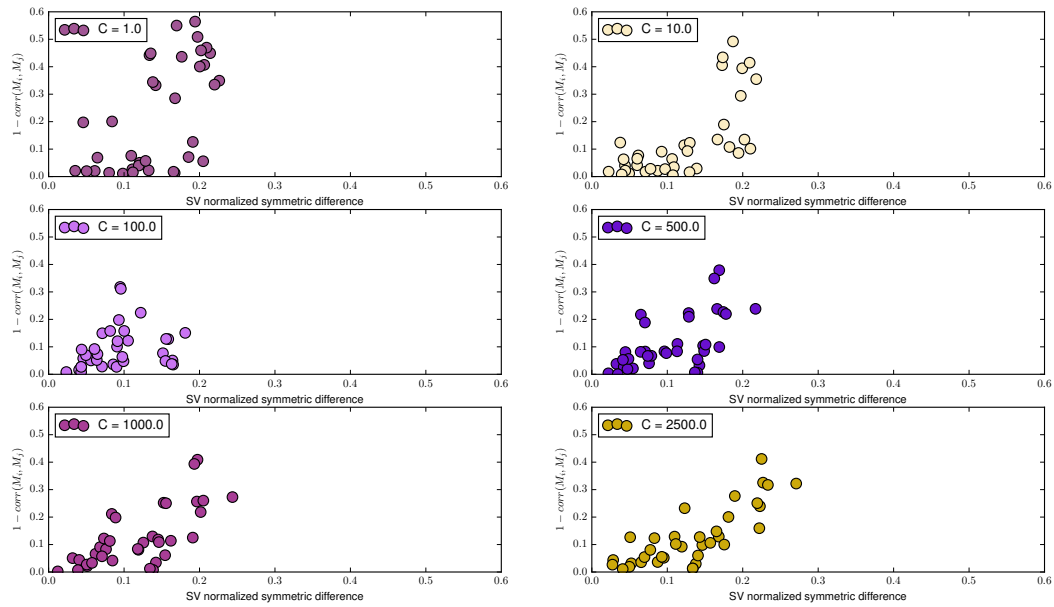


Рис. 2. Wine quality

Таблица 2. Wine info

	$\text{mean}(\#SV)$	$\text{mean}(1 - \text{corr}(M_i, M_j))$	$\text{mean}(\rho_{X^l}(K_i, K_j))$	Correlation
$C = 1.0$	3284.1	0.220	0.144	0.600
$C = 10.0$	3284.9	0.130	0.121	0.687
$C = 100.0$	3275.0	0.091	0.091	0.270
$C = 500.0$	3252.6	0.110	0.105	0.591
$C = 1000.0$	3235.2	0.124	0.118	0.694
$C = 2500.0$	3208.6	0.127	0.133	0.795

Heart dataset

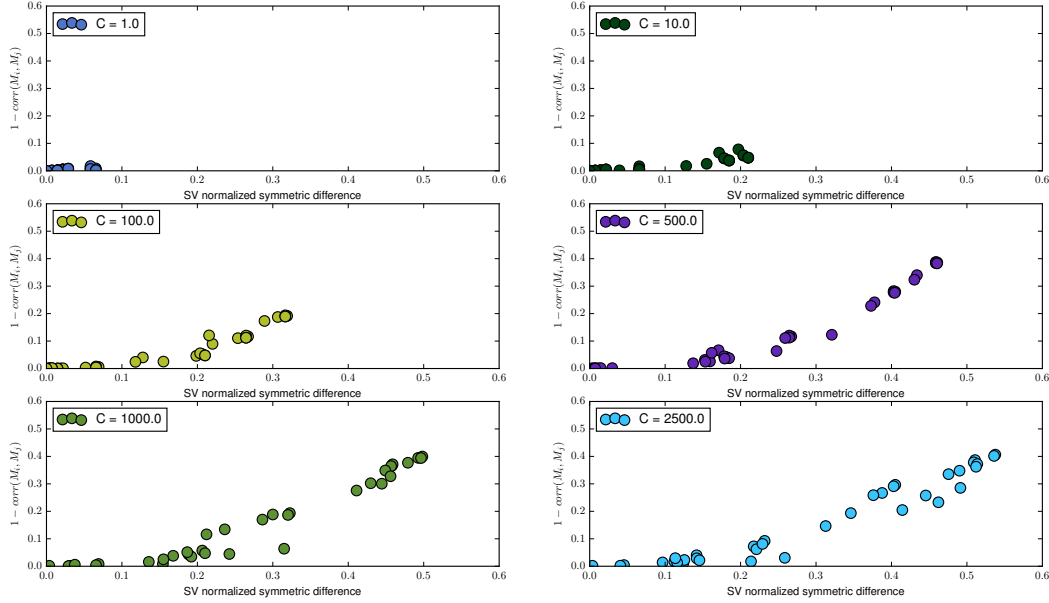


Рис. 3. Heart disease

Таблица 3. Heart info

	mean(#SV)	mean($1 - \text{corr}(M_i, M_j)$)	mean($\rho_{X^l}(K_i, K_j)$)	Correlation
$C = 1.0$	272.0	0.003	0.027	0.608
$C = 10.0$	260.8	0.020	0.088	0.929
$C = 100.0$	249.1	0.063	0.152	0.927
$C = 500.0$	231.9	0.135	0.238	0.940
$C = 1000.0$	223.1	0.157	0.268	0.953
$C = 2500.0$	211.4	0.166	0.297	0.962

Общие выводы

Уже по этим данным можно сказать, что (будем называть $1 - \text{corr}(M_i, M_j)$ *расстоянием между отступами*).

- С ростом константы регуляризации расстояние между ядрами и расстояние между их отступами лучше коррелируют между собой.
- При высоких параметре регуляризации коэффициент корреляции Пирсона достигает более 0.8, то есть расстояния практически линейно зависят друг от друга.
- Вектора средних ядерных и отступных расстояний коррелируют по-разному на различных датасетах (на Wine и Heart корреляции Пирсона 0.85 и 0.99 соответственно, на German — -0.92).

Литература

- [1] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.
- [2] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, 1992.
- [3] Konstantin Vyacheslavovich Vorontsov. Mathematical methods of learning from examples (machine learning theory).
- [4] Alex J Smola, Bernhard Sch, and B Schölkopf. A Tutorial on Support Vector Regression. *Statistics and Computing*, 14(3):199–222, 2004.
- [5] Leo Breiman. Bagging Predictors. *Machine Learning*, 24(421):123–140, 1996.
- [6] Y. Freund. Boosting a Weak Learning Algorithm by Majority, 1995.
- [7] J. Franke and E. Mandler. A comparison of two approaches for combining the votes of cooperating classifiers. *Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems*, pages 1–4, 1992.
- [8] Tin Kam T.K. Ho, JJ Hull, Sargur N SN Srihari, and Senior Member. Decision combination in multiple classifier systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(1):66–75, 1994.
- [9] J. Kittler, M. Hater, and R. P W Duin. Combining classifiers. *Proceedings - International Conference on Pattern Recognition*, 2(3):897–901, 1996.
- [10] Manuel Martin-merino and Manuel Mart. Combining SVM Classifiers for Email Anti-spam Filtering. 4507(February), 2007.
- [11] D. Gorgevik and D. Cakmakov. Handwritten Digit Recognition by Combining SVM Classifiers. *EUROCON 2005 - The International Conference on Computer as a Tool*, 2(February):1393–1396, 2005.
- [12] Martin Dyrba, Michel Grothe, Thomas Kirste, and Stefan J. Teipel. Multimodal analysis of functional and structural disconnection in Alzheimer’s disease using multiple kernel SVM. *Human Brain Mapping*, 36(6):2118–2131, 2015.
- [13] S.S. Bucak, R. Jin, and Ak. Jain. Multiple Kernel Learning for Visual Object Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1354–1369, 2014.
- [14] Salah Althloothi, Mohammad H. Mahoor, Xiao Zhang, and Richard M. Voyles. Human activity recognition using multi-features and multiple kernel learning. *Pattern Recognition*, 47(5):1800–1812, 2014.
- [15] Stefan Aeberhard. Wine data set, 1991.
- [16] Dr. Hans Hofmann. Statlog (german credit data) data set, 1994.
- [17] Andras Janosi M.D, M.D William Steinbrunn, M.D Matthias Pfisterer, and Ph.D Robert Detrano, M.D. Heart disease data set, 1988.