FRI-ONLINE-1-CCT1-03

A REVIEW ON THE MODIFICATIONS OF THE K-NEAREST NEIGHBOR ALGORITHM[3]

Tsvetelina Mladenova
Department of Computer Systems and Technologies
"Angel Kanchev" Univesity of Ruse
E-mail: tsmladenova@uni-ruse.bg

Irena Valova
Department of Computer Systems and Technologies
"Angel Kanchev" Univesity of Ruse
E-mail: ivalova@ecs.uni-ruse.bg

Abstract: *The implementation of machine learning in businesses, and in particular small and medium businesses, depends to a large extent on the ability of managers and decision-makers to understand the used algorithm. Faced with the need to make a decision concerning their company, many managers are sceptical to trust a system they do not understand. That is why one of the keys for the successful implementation of machine learning in business is the interpretability factor.*

This publication reviews one of the most popular, easy-to-understand, and preferred machine learning algorithms - the K-Nearest Neighbor. Although the algorithm has proven to be useful and accurate, it has some serious drawbacks and shortcomings that many authors have explored and researched. This publication is a study of some of the most significant modifications of the traditional KNN algorithm and the problems they solve. The reviewed modifications are classified by:

- *The problem they are solving*
- *The main idea of the modification*
- *Positive characteristics of the modification*
- *Negative characteristics of the modification*
- *Area of application*

Keywords: *Machine learning, K-nearest neighbor, KNN*

ВЪВЕДЕНИЕ

Методът на К-най-близките съседи (KNN), на англ. K-Nearest Neighbors, е един от най-разпространените и използвани алгоритми. Алгоритъмът намира приложение в решаването на класификационни проблеми, а лесната му имплементацията и сравнително простите операции, го правят предпочитан избор, когато данните не са особено големи и има необходимост от задълбочено познание за това как алгоритъмът е стигнал до дадено решение.

Основните стъпки при класификацията с KNN са три, фиг 1. Първата стъпка при класификацията е изчислението на разстоянията от некласифицирания обект до всеки обект от обучителното множество. Тъй като тази операция се извършва за всеки некласифициран обект, понеже местоположението на всеки обект е различно, времето за класификация е голямо. Това е един от недостатъците на алгоритъма. Най-често, методът за изчисление на разстоянията е Евклидовото разстояние, като някои други методи са: разстояние на Минковски, Манхатаново разстояние.

Вторият етап е определянето на съседството. На този етап изчислените разстояния се обработват и се избират k съседите, които ще образуват съседството. Крайната класификация на непознатия обект силно зависи от съседите, които са избрани на този етап. Един от

[3] Докладът е представен на заседание на секция 3.2 на 29 октомври 2021 с оригинално заглавие A REVIEW ON THE MODIFICATIONS OF THE K-NEAREST NEIGHBOR ALGORITHM

използваните методи се състои във възходяща подредба на съседите в зависимост от изчисленото разстояние.

Третият, и последен, етап е определянето на класа към който принадлежи непознатият обект. Този етап силно зависи от определеното съседство. Именно съседите в него, определят към кой клас ще принадлежи непознатия обект. Най-използван е методът за правилото на мнозинството, при който се разглежда броя на представителите на класовете. Класът, имащ най-голям брой представители в съседството, е класът към който бива класифициран обектът.

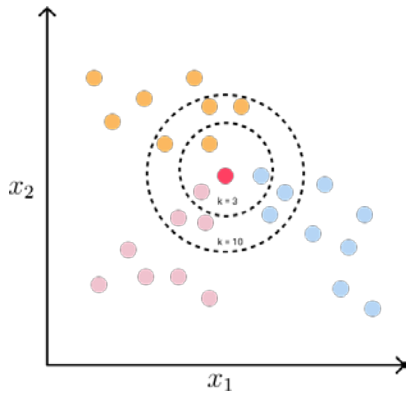
Фиг. 1 Стъпки при класификация с KNN

ПРОБЛЕМИ ПРИ КЛАСИФИКАЦИЯТА ПО МЕТОДА НА НАЙ-БЛИЗКИТЕ СЪСЕДИ

Модификациите на алгоритъма са в посока справяне с някои от проблемите, характерни за него. Ali, P. U. S. and Ventakeswaran, D. C. J. (2011) предлагат алгоритъм, базиран на теорията за доказателствата, Gates, G. (1972) предлага редуциран най-близък съсед, който обръща внимание както на избора на съседите, така и на правилото за класификация. Guo, G. *et al.* (2003) предлагат подход, базиран на модели на съседите, а Keller, J. M., Gray, M. R. and Givens, J. A. (1985) разглеждат алгоритъма като размит (на англ. *fuzzy*). Други модификации, заслужаващи да бъдат споменати са Li, S. Z., Chan, K. L. and Wang, C. (2000), Mitani, Y. and Hamamoto, Y. (2006), Tao, Y., Papadias, D. and Lian, X. (2004), Wang, H. and Bell, D. (2004), Zeng, Y., Yang, Y. and Zhao, L. (2009), Zhang, M.-L. and Zhou, Z.-H. (2007).

A) Чувствителност на k параметъра

Стойността на k параметъра определя броя на съседите, които ще влязат в съседството. То, от своя страна, е определящо за крайното решение на класификатора. Следователно, можем да твърдим, че правилното определяне на k стойността има пряко значение върху точността на алгоритъма. Да приемем, че за множеството от обекти, визуализирано на фиг. 2, изберем $k = 3$. В съседството ще влязат по един представител на всеки клас, което ще намали възможността за правилна класификация. От друга страна, ако за стойност на k изберем 10, то разпределението на съседите ще е в полза на един от класовете, въпреки, че не е изключено некласифицирания обект да се намира по-близо до някои от класовете с по-малко представителство.



Фиг. 2 Избор на k стойност

При избора на стойност за k параметъра, трябва да се има предвид:

- Размерът на обучителното множество;
- Разпределението на обектите в множеството;
- Броят на уникалните класове в обучителното множество;

В) Избор на съседство

Изчислението на разстоянието между обектите в множеството и некласифицираните обекти, най-често става с Евклидовото разстояние, но в зависимост от данните в множеството, наличието на аномалии, размера на множеството, както и други характеристики, може да се окаже, че Евклидовото разстояние не дава добри резултати. В такива случаи се налага използването на друга метрика за изчисление. Тъй като вариантите за изчисление на разстоянията не са краен брой, това понякога води до затруднения в използването на метода на най-близките съседи. Табл. 1 представя някои от използваните метрики за изчисление.

Табл. 1 Формули за изчисление на разстоянието

Метод	**Формула**	**Пояснение**
Разстояние на Минковски (Minkowski)	$\sqrt[P]{\sum \lvert x_i - y_i \rvert^P}$	x_i, y_i – точки от множеството p - норма
Манхатаново разстояние	$\sum \lvert x_i - y_i \rvert$	x_i, y_i – точки от множеството
Евклидово разстояние	$\sqrt{\sum \lvert x_i - y_i \rvert^2}$	x_i, y_i – точки от множеството
Чебишев (Chebychev)	$max \lvert x_i - y_i \rvert$	x_i, y_i – точки от множеството
Хаминг (Hamming)	$\sum x_i \oplus y_i$	x_i, y_i – точки от множеството
Жакард (Jaccard)	$1 - \frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$	A, B – множества

Освен изчислението на съседството, алгоритъмът трябва да „реши“ кои обекти ще влязат в съседството. През 1996г. (Chaudhuri, B. B., 1996) предлагат т.нар. „симетрично съседство“, на англ. Nearest Centroid Neighborhood (NCN), което се базира на идеята, че съседите не трябва

да са просто най-близките съседи, а че трябва да са избрани така, че да образуват симетрично съседство около некласифицирания обект. Така, съседи, които биха били избрани по традиционния начин – само с изчисление на разстоянията, могат да бъдат пренебрегнати в полза на по-отдалечени, но по-симетрични съседи. В Табл. 2 е представен псевдо-кода на алгоритъма.

Табл. 2 Псевдо-код на алгоритъма NCN

Псевдо-код на алгоритъма Nearest Centroid Neighborhood

Вход: (X, , k, p, Dist)

Изход: r

$Q \leftarrow \emptyset$
$q_1 \leftarrow findNN(X, p)$ — q_1 е най-близкия съсед
$Q \leftarrow q_1$
$X_{aux} \leftarrow X - \{q_1\}$
$r \leftarrow Dist(q_1, p)$
$j \leftarrow 1$
Докато $j \leq k$:
    $j \leftarrow j + 1$
    $dist_{min} \leftarrow \infty$
    За всяка точка $x_i \in X_{aux}$:
        $M \leftarrow computeCentroid(Q \cup \{x_i\})$
        Ако $Dist(M, p) \leq dist_min$:
            $q_j \leftarrow x_i$
            $dist_min \leftarrow Dist(M, p)$ — „обновяване“ на минималното разстояние
    $Q \leftarrow Q \cup \{q_j\}$
    $r \leftarrow \max(r, Dist(q_j, p))$ — „обновяване“ на радиуса, при необходимост
    $X_{aux} \leftarrow X_{aux} - \{q_j\}$

C) Правило за класификация

Най-често използваното правило за класификация е правилото на множеството. И макар, да се е доказало като ефективно работещ вариант, има случаи, в които не дава оптимални резултати. Доказано, то не дава добри резултати, когато границите между класовете са размити, множеството е небалансирано (Liu, W. and Chawla, S., 2011) или е голямо и сбито (Liu, H. and Zhang, S., 2012).

Алтернатива на правилото на множеството е правилото на претегленото множество (Dudani, S. A., 1976). Претегленото множество добавя стъпка на изчисление на теглата на всеки съсед в съседството, като след това сумира всички тегла. Класът, чиито сумирани тегла са „най-тежки“ е класът, към който ще бъде класифициран обекта.

Теглата могат да бъдат изчислени по някоя от формулите, представени в табл. 3.

Табл. 3 Формули за изчисление на теглото

Формула за изчисление на теглото	**Където:**
$w_i = \begin{cases} \frac{d_k - d_i}{d_k - d_1}, & d_k \neq d_1 \\ 1, & d_k = d_1 \end{cases}$	d_k – разстоянието до последния съсед d_i – разстоянието до i-я съсед d_1 – разстоянието до първия съсед
$w_i = \begin{cases} \frac{d_k - d_i}{d_k - d_1} \times \frac{d_k + d_1}{d_k + d_i}, & d_k \neq d_1 \\ 1, & d_k = d_1 \end{cases}$	d_k – разстоянието до последния съсед d_i – разстоянието до i-я съсед d_1 – разстоянието до първия съсед
$w_i = \frac{d_k - d_i}{d_k - d_1} * \frac{1}{i}$	d_k – разстоянието до последния съсед d_i – разстоянието до i-я съсед

Формула за изчисление на теглото	Където:
	d_1 – разстоянието до първия съсед i – i-я елемент от съседството

ИЗВОДИ

Най-често използван метод за изчисление на k стойността е експерименталния - многократно обучение на модел и избор на стойността, при която ефективността на модела започва да намалява. Обикновено, избраната стойност за к зависи от обучителното множество.

Най-използвания метод е Евклидовото разстояние. Въпреки, че има множество други методи за изчисление на разстоянието, Евклидовото разстояние е доказано, един от най-ефективните методи. Наличието на алгоритми, които позволяват по-гъвкав избор на съседите дава възможност за по-голяма точност при класификацията.

Правилото на мнозинството е методът за вземане на решение, който доминира. Въпреки това, има и други алтернативи, които са подходящи в частните случаи, когато традиционното правило не би могло да даде добри класификационни резултати.

REFERENCES

Ali, P. U. S. and Ventakeswaran, D. C. J. (2011) ‘Improved Evidence Theoretic kNN Classifier based on Theory of Evidence’, *International Journal of Computer Applications*. Citeseer.

Chaudhuri, B. B. (1996) ‘A new definition of neighborhood of a point in multi-dimensional space’, *Pattern Recognition Letters*. Elsevier.

Dudani, S. A. (1976) ‘The distance-weighted k-nearest-neighbor rule’, *IEEE Transactions on Systems, Man, and Cybernetics*. IEEE.

Gates, G. (1972) ‘The reduced nearest neighbor rule (corresp.)’, *IEEE transactions on information theory*. IEEE.

Guo, G. *et al.* (2003) ‘KNN model-based approach in classification’, *OTM Confederated International Conferences*.

Keller, J. M., Gray, M. R. and Givens, J. A. (1985) ‘A fuzzy k-nearest neighbor algorithm’, *IEEE transactions on systems, man, and cybernetics*. IEEE.

Korn, F. and Muthukrishnan, S. (2000) ‘Influence sets based on reverse nearest neighbor queries’, *ACM Sigmod Record*. ACM New York, NY, USA.

Li, S. Z., Chan, K. L. and Wang, C. (2000) ‘Performance evaluation of the nearest feature line method in image classification and retrieval’, *IEEE Transactions on Pattern Analysis and Machine Intelligence*. IEEE.

Liu, W. and Chawla, S. (2011) ‘Class confidence weighted knn algorithms for imbalanced data sets’, *Pacific-Asia conference on knowledge discovery and data mining*.

Liu, H. and Zhang, S. (2012) ‘Noisy data elimination using mutual k-nearest neighbor for classification mining’, *Journal of Systems and Software*. Elsevier.

Mitani, Y. and Hamamoto, Y. (2006) ‘A local mean-based nonparametric classifier’, *Pattern Recognition Letters*. Elsevier.

Tao, Y., Papadias, D. and Lian, X. (2004) ‘Reverse knn search in arbitrary dimensionality’, *Proceedings of the Very Large Data Bases Conference (VLDB), Toronto*.

Wang, H. and Bell, D. (2004) ‘Extended k-nearest neighbours based on evidence theory’, *The Computer Journal*. OUP.

Zeng, Y., Yang, Y. and Zhao, L. (2009) ‘Pseudo nearest neighbor rule for pattern classification’, *Expert Systems with Applications*. Elsevier.

Zhang, M.-L. and Zhou, Z.-H. (2007) ‘ML-KNN: A lazy learning approach to multi-label learning’, *Pattern recognition*. Elsevier.