Pattern Recognition 47 (2014) 2702–2720

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

LiNearN: A new approach to nearest neighbour density estimator

Jonathan R. Wells [a,*], Kai Ming Ting [b], Takashi Washio [c]

[a] Faculty of Information Technology, Monash University, Australia
[b] School of Information Technology, Federation University, Australia
[c] The Institute of Scientific and Industrial Research, Osaka University, Japan

ARTICLE INFO

Article history:
Received 30 January 2013
Received in revised form
9 September 2013
Accepted 23 January 2014
Available online 1 February 2014

Keywords:
k-nearest neighbour
Density-based
Anomaly detection
Clustering

ABSTRACT

Despite their wide spread use, nearest neighbour density estimators have two fundamental limitations: $O(n^2)$ time complexity and $O(n)$ space complexity. Both limitations constrain nearest neighbour density estimators to small data sets only. Recent progress using indexing schemes has improved to near linear time complexity only.

We propose a new approach, called *LiNearN* for *Li*near time *Near*est *N*eighbour algorithm, that yields the first nearest neighbour density estimator having $O(n)$ time complexity and constant space complexity, as far as we know. This is achieved without using any indexing scheme because LiNearN uses a subsampling approach for which the subsample values are significantly less than the data size. Like existing density estimators, our asymptotic analysis reveals that the new density estimator has a parameter to trade off between bias and variance. We show that algorithms based on the new nearest neighbour density estimator can easily scale up to data sets with millions of instances in anomaly detection and clustering tasks.

1. Introduction and motivation

Existing methods have utilised nearest neighbour density estimators as the basis to solve all facets of pattern recognition problems from classification and regression to clustering and anomaly detection [13,5,15,11,7].

While existing nearest neighbour density estimators have been effective, the time complexity is still basically $O(n^2)$ because of the need to find the nearest neighbour for every instance in a given data set. This makes existing methods utilising nearest neighbour density estimator impractical for problems with large data sets. Recent research has substantially improved the *k*-nearest neighbour search by introducing various indexing schemes to speed up the search (e.g., Cover Trees [9], M-Trees [12] and R*-Tree [8]) to near linear time complexity.

The premise of the current research is that finding the nearest neighbour for every instance in the given data set is inevitable which leads to $O(n^2)$ time complexity. Since the aim is to do density estimation, we reject this premise and find a way to reduce the number of pair-wise distance calculations required to achieve this aim.

We propose a new approach to nearest neighbour density estimation. Instead of focusing on speeding up the nearest neighbour search, the new approach first generates many local regions from subsamples and then produces the final result in an ensemble method. The speedup is achieved because the size of the subsamples required is significantly smaller than the given data set. This not only eliminates the need of using an indexing scheme but enables the new density estimator to run in orders of magnitude faster than existing nearest neighbour density estimators.

We make three contributions in this paper:

1. Introduce a new nearest neighbour density estimator that defines local neighbourhoods using nearest neighbours in each of the many subsamples by building a region centered at each instance. This differs from the existing nearest neighbour density estimators where the local neighbourhoods are defined based on either *k* nearest neighbours or a fixed radius.
2. Provide an asymptotic analysis and it reveals that the new density estimator has a parameter which trades off between bias and variance, as in existing density estimators such as *k*-nearest neighbour density estimators.
3. Demonstrate the advantages of the new approach over the existing nearest neighbour density estimators in two tasks: anomaly detection and clustering. The new approach reduces the time complexity from $O(n^2)$ to $O(n)$ and the space complexity complexity from $O(n)$ to a constant. We call the new approach *LiNearN* for *Li*near time *Near*est *N*eighbour algorithm.

Since nearest neighbour density estimators are the core mechanism in many pattern recognition algorithms, we will begin the next section with a description of existing nearest neighbour

* Corresponding author.
E-mail addresses: jonathan.wells@monash.edu (J.R. Wells), kaiming.ting@federation.edu.au (K.M. Ting), washio@ar.sanken.osaka-u.ac.jp (T. Washio).

http://dx.doi.org/10.1016/j.patcog.2014.01.013
0031-3203

density estimators. Section 3 introduces the new nearest neighbour density estimator and provides the asymptotic analysis. Section 4 describes how both the existing and the new nearest neighbour density estimators are applied to anomaly detection and clustering tasks. Section 5 reports the empirical evaluation results. Discussion and the conclusions are provided in the last two sections.

2. Existing nearest neighbour density estimators

We describe three existing nearest neighbour density estimators below.

1. A k-nearest neighbour (k-NN) density estimator can be expressed as follows [32,11]:

$$f_{kNN}(x) = \frac{|N(x,k)|}{n \sum_{x' \in N(x,k)} \|x - x'\|_p}$$

where $N(x,k)$ is the set of k-nearest neighbours to x; and $|S|$ denotes the cardinality of set S, and $\|x-x'\|_p$ denotes the distance measured by L^p-norm between x and x'. The search for nearest neighbours is conducted over D of size n, by using the L^p-norm distance where D is the given data set.

2. A kth nearest neighbour density estimator is defined as follows [25]:

$$f_{k^{th}NN}(x) = \frac{|N(x,k)|}{n\alpha(d,p)\|x - x_k\|_p^d}$$

where x_k is the kth nearest neighbour to x and $\alpha(d,p)$ is the volume of an unit ball in $(\Re^d, L^p)$

3. An ε-neighbourhood density estimator is defined as follows:

$$f_{\varepsilon}(x) = \frac{|N_{\varepsilon}(x)|}{n\varepsilon}$$

where $N_{\varepsilon}(x) = \{q \in D \mid \|x-q\|_p \leq \varepsilon\}$. Since the denominator $n\varepsilon$ is the same for all x, it is usually omitted in the implementation (e.g., in DBSCAN [15]).

Each of the above determines a local neighbourhood based on a global parameter, i.e., k or ε; and the density is calculated based on one variable: distance of k-nearest neighbours in f_{kNN} or $f_{k^{th}NN}$ since the numerator is a constant k; and $N_{\varepsilon}(\cdot)$ in f_{ε} since the denominator is a constant.

In addition, the nearest neighbour search is conducted over the entire data set, D, which is the main computational expense of the whole process; therefore, leading to a time complexity of $O(n^2)$ for n queries. Research has focused on reducing this cost by devising different indexing schemes.

We suggest a new approach to compute density based on nearest neighbour with the following distinguishing features:

- Both the number of instances in the local neighbourhood and its volume are adaptive to the data distribution in the local region; neither is fixed by a global parameter, unlike $f_{kNN}(\cdot)$, $f_{k^{th}NN}(\cdot)$ and $f_{\varepsilon}(\cdot)$.
- The nearest neighbour search is conducted over a data subset which is significantly smaller than the given data set.

We describe the new density estimator in the next section.

3. New nearest neighbour density estimator

We propose a new nearest neighbour density estimator, called *LiNearN* for *Li*near time *Near*est *N*eighbour algorithm. It estimates the density for a point x by averaging densities of multiple local regions covering x. Whilst the local regions could be implemented in different ways, we focus on deriving the local regions using nearest neighbours. Because these local regions can be defined by using a significantly smaller data set than the given data set, the computational expense for nearest neighbour search is reduced to such an extent that an indexing scheme becomes unnecessary.

We describe LiNearN in the following five subsections. After describing the key differences between the new and existing density estimators in the first subsection, LiNearN is formally defined in the second subsection with an illustration in the third subsection. The asymptotic error analysis is given in the fourth subsection followed by its implementation in the fifth subsection.

3.1. Key differences

The key differences between LiNearN and existing density estimators based on the nearest neighbour are shown in Table 1. Since all parameters, except n, are constant and both $\psi \ll n$ and $\Psi \ll n$ (see definitions in Section 3.2), the time complexity of LiNearN is $O(n)$, which is significantly smaller than $O(n^2)$ or $O(n \log n)$.

Unlike ε-neighbourhood density estimator which employs a global ε (where every local region has the same size), LiNearN adapts the size of each local region to the local data distribution. For example, sparse regions have large local regions, whereas dense regions have small local regions. While k-nearest neighbour density estimator can adapt to local data distributions in simple

Table 1
Key differences between existing nearest neighbour algorithms and LiNearN in terms of methodology and time complexity.

	Existing NN	LiNearN
Methodology	Single model Density for each $x \in D$ is derived from a single local region via NN searches (e.g., f_{kNN}, $f_{k^{th}NN}$ or f_{ε}) Indexing[a] is required to speed up NN search. Often rely on triangle inequality to prune the search space	Multiple models Density for each $x \in D$ is derived from many local regions (LR) NN search without indexing 1. NN search in a subset of D (t times) to define LR 2. NN search to make the final estimation for each $x \in D$
Time complexity		
LR building	Not applicable	1. $\psi(\psi+\Psi)t$
Index building	Nil or $n \log n$[b]	Not applicable
n queries	n^2 or $n \log n$	2. ψnt

[a] An alternative to indexing is clustering based search [27] which often needs higher time cost than indexing.
[b] Without indexing, n queries in existing nearest neighbour algorithms have $O(n^2)$ time complexity; with indexing methods such as Cover Trees [9] and M-Trees [12], n queries have $O(n \log n)$ time complexity.