# kdANN+: A Rapid AkNN Classifier for Big Data

Nikolaos Nodarakis[1], Evaggelia Pitoura[2], Spyros Sioutas[3], Athanasios Tsakalidis[1], Dimitrios Tsoumakos[3], and Giannis Tzimas[4]

[1] Computer Engineering and Informatics Department, University of Patras, 26500 Patras, Greece
{nodarakis,tsak}@ceid.upatras.gr
[2] Computer Science Department, University of Ioannina
pitoura@cs.uoi.gr
[3] Department of Informatics, Ionian University, 49100 Corfu, Greece
{sioutas,dtsouma}@ionio.gr
[4] Computer & Informatics Engineering Department, Technological Educational Institute of Western Greece, 26334 Patras, Greece
tzimas@cti.gr

**Abstract.** A $k$-nearest neighbor ($k$NN) query determines the $k$ nearest points, using distance metrics, from a given location. An all $k$-nearest neighbor (A$k$NN) query constitutes a variation of a $k$NN query and retrieves the $k$ nearest points for each point inside a database. Their main usage resonates in spatial databases and they consist the backbone of many location-based applications and not only. In this work, we propose a novel method for classifying multidimensional data using an A$k$NN algorithm in the MapReduce framework. Our approach exploits space decomposition techniques for processing the classification procedure in a parallel and distributed manner. To our knowledge, we are the first to study the $k$NN classification of multidimensional objects under this perspective. Through an extensive experimental evaluation we prove that our solution is efficient, robust and scalable in processing the given queries.

**Keywords:** classification· nearest neighbor· MapReduce· Hadoop· multidimensional data· query processing

## 1 Introduction

Classification is the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. One of the algorithms for data classification uses the $k$NN approach [10]. It computes the $k$ nearest neighbors (belonging to the training dataset) of a new object and classifies it to the category that belongs the majority of its neighbors.

A $k$-nearest neighbor query [19] computes the $k$ nearest points, using distance metrics, from a specific location and is an operation that is widely used in spatial databases. An all $k$-nearest neighbor query constitutes a variation of a $k$NN query and retrieves the $k$ nearest points for each point inside a dataset in a single query process. There is a wide diversity of applications that A$k$NN queries can be harnessed. The classification problem is one of them. Furthermore, they are widely used by location based services [13]. For example, consider users that send their location to a web server to process a request using a position anonymization system in order to protect their privacy from insidious acts. This anonymization system may use an A$k$NN algorithm to calculate the $k$ nearest neighbors for each user. After that, it sends to the server the locations of the neighbors along with the location of the user that made the request at the first place. In addition, many algorithms have been developed to optimize and speed up the join process in databases using the $k$NN approach.

Although A$k$NN is a fundamental query type, it is computationally very expensive. The naive approach is to search for every point the whole dataset in order to estimate its $k$-NN list. This leads to an $O\left(n^2\right)$ time complexity assuming that $n$ is the cardinality of the dataset. As a result, quite a few centralized algorithms and structures (M-trees, R-trees, space-filling curves, etc.) have been developed towards this direction [6, 12, 15, 31]. However, as the volume of datasets grows rapidly even these algorithms cannot cope with the computational burden produced by an A$k$NN query process. Consequently, high scalable implementations are required. Cloud computing technologies provide tools and infrastructure to create such solutions and manage the input data in a distributed way among multiple servers. The most popular and notably efficient tool is the *MapReduce* [9] programming model, developed by Google, for processing large-scale data.

In this paper, we propose a method for efficient multidimensional data classification using A$k$NN queries in a single batch-based process in *Hadoop* [22, 25], the open source MapReduce implementation. The basic idea is to decompose the space, where the data belongs, into smaller partitions. Afterwards, we get the $k$ nearest neighbors for each point to be classified only by searching the appropriate partitions. Finally, we add it to the category it belongs based on the class that the majority of its neighbors belongs. The space decomposition relies on the data distribution of the training dataset.

More specifically, we sum up the technical contributions of our paper as follows:

- We present an implementation of a classification algorithm based on A$k$NN queries using MapReduce. We apply space decomposition techniques (based on data distribution) that divides the data into smaller groups. For each point we search for candidate $k$-NN objects only in a few groups. The granularity of the decomposition is a key factor for the performance of the algorithm and we analyze it further in Section 6.1. At first, the algorithm defines a search area for each point and investigates for $k$-NN points in the groups covered by this area. If the search area of a point does not include at least $k$ neighbors, it is gradually expanded until the desired number is

reached. Finally, we classify the point to the category that belongs the majority of its neighbors. The implementation defines the MapReduce jobs with no modifications to the original Hadoop framework.
- We provide an extension for $d > 3$ in Section 5 ($d$ stands for dimensionality).
- We evaluate our solution through an experimental evaluation against large scale data up to 4 dimensions. Furthermore, we study various parameters that can affect the total computational cost of our method using real and synthetic datasets. The results prove that our solution is efficient, robust and scalable.

The rest of the paper is organized as follows: Section 2 discusses related work. Section 3 presents the initial idea of the algorithm, our technical contributions and some examples of how the algorithm works. Section 4 presents a detailed analysis of the classification process developed in Hadoop. Section 5 provides an extension for $d > 3$ and Section 6 presents the experiments that where conducted in the context of this work. Finally, Section 7 concludes the paper and Section 8 presents future steps.

## 2 Related Work

A$k$NN queries have been extensively studied in literature. In [15], a method based on M-trees is proposed that processes A$k$NN spatial network queries. The experimental evaluation runs over a road network dataset for small $k$ values. In addition, a structure that is popular for answering efficiently to $k$NN queries is R-tree [19]. Assuming that we execute a $k$NN query for all elements stored in the R-tree, we facilitate the A$k$NN query process with such indexes. Pruning techniques can be combined with such structures to deliver better results [6, 12]. Mobile networks are also a domain where A$k$NN find application as shown in [4]. Their work suggest a centralized algorithm that identifies to every smartphone user its $k$ geographically nearest neighbors in $O(n \cdot (k + l))$ time, where $n$ denotes the number of users and $l$ is a network-specific parameter. Moreover, efforts have been made to design low computational cost methods that execute such queries in spatial databases. For instance, [27] studies both the $k$NN query and the $k$NN join in a relational database. Their approach guarantees to find the approximate $k$NN with only logarithmic number of page accesses in expectation with a constant approximation ratio. Also, it can be extended to find the exact $k$NN efficiently in any fixed dimension. The works in [26, 29] propose algorithms to answer $k$NN join.

The methods proposed above can handle data of small size in one or more dimensions, thus their use is limited in centralized environments only. During the recent years, the researchers have focused on developing approaches that are applicable in distributed environments, like our method, and can manipulate big data in an efficient manner. The MapReduce framework seems to be suitable for processing such queries. For example, in [28] the discussed approach splits the target space in smaller cells and looks into appropriate cells where $k$-NN objects are located, but applies only in 2-dimensional data. Our method speeds up the

naive solution of [28] by eliminating the merging step, as it is a major drawback. We have to denote here that in [28] it is claimed that the computation of the merging step can be performed in one node since we just consider statistic values. But this is not entirely true since this process can derive a notable computational burden as we increase dimensions and/or data size, something that is confirmed in the experimental evaluation. In addition, the merging step can produce sizeable groups of points, especially as $k$ increments, that can overload the first step of the A$k$NN process. Moreover, our method applies for more dimensions. Especially, for $d >= 3$ the multidimensional extension is not straightforward at all.

In [21], locality sensitive hashing (LSH) is used together with a MapReduce implementation for processing $k$NN queries over large multidimensional datasets. This solution suggests an approximate algorithm like the work in [30] (H-zkNNJ) but we focus on exact processing of A$k$NN queries. Furthermore, A$k$NN queries are utilized along with MapReduce to speed up and optimize the join process over different datasets [2, 17] or support non-equi joins [24]. Moreover, [3] makes use of a R-tree based method to process $k$NN joins efficiently. Together with $k$NN, many other popular spatial queries have been studied and implemented efficiently on top of Hadoop/HBase frameworks [1, 11, 16].

In [5] a minimum spanning tree based classification model is introduced and it can be viewed as an intermediate model between the traditional $k$-nearest neighbor method and cluster based classification method. Another approach presented in [14] recommends parallel implementation methods of several classification algorithms, including $k$-nearest neighbor, bayesian model, decision tree. However, it does not contemplate neither the perspective of dimensionality nor parameter $k$.

In brief, our proposed method implemented in the Hadoop MapReduce framework, extends the traditional $k$NN classification algorithm and processes exact A$k$NN queries over massive multidimensional data. In this way, we achieve to classify a huge amount of objects in a single batch-based process. Compared to the aforementioned solutions, our method does not focus solely on the join operator but provides a more generalized framework to process A$k$NN queries. In other words, we boost the performance of the A$k$NN query process regardless the context of use of the query ($k$NN join, A$k$NN classification, etc.) The experimental evaluation considers a wide diversity of factors that can affect the execution time such as the value of $k$, the granularity of space decomposition, dimensionality and data distribution.

## 3 Overview of Classification Algorithm

In this section, we first define some notation and provide some definitions used throughout this paper. Table 1 lists the symbols and their meanings. Next, we outline the architecture of MapReduce model. Finally, we give a brief review of the method our solution relies on and then we extend it for more dimensions and tackle some performance issues.

**Table 1.** Symbols and their meanings

| | |
|---|---|
| $n$ | granularity of space decomposition |
| $k$ | number of nearest neighbors |
| $d$ | dimensionality |
| $D$ | a $d$-dimensional metric space |
| $dist(r, s)$ | the distance from $r$ to $s$ |
| $kNN(r, S)$ | the $k$ nearest neighbors of $r$ from $S$ |
| $AkNNC(R, S)$ | $\forall r \in R$ classify $r$ based on $kNN(r, S)$ |
| $ICCH$ | interval, cell cube or hypercube |
| $ICSH$ | interval, circle, sphere or hypersphere |
| $I$ | input dataset |
| $T$ | training dataset |
| $c_r$ | the class of point $r$ |
| $C_T$ | the set of classes of dataset $T$ |
| $S_I$ | size of input dataset |
| $S_T$ | size of training dataset |
| $M$ | total number of Map tasks |
| $R$ | total number of Reduce tasks |

### 3.1 Definitions

We consider points in a $d$-dimensional metric space $D$. Given two points $r$ and $s$ we define as $dist(r, s)$ the distance between $r$ and $s$ in $D$. In this paper, we used the distance measure of Euclidean distance

$$(r, s) = \sqrt{\sum_{i=1}^{d} (r[i] - s[i])^2}$$

where $r[i]$ (respectively $s[i]$) denote the value of $r$ (respectively $s$) along the $i$-th dimension in $D$. Without loss of generality, alternative distance measures (i.e. Manhattan distance) can be applied to our solution.

**Definition 1.** ***kNN:*** *Given a point $r$, a dataset $S$ and an integer $k$, the $k$ nearest neighbors of $r$ from $S$, denoted as $kNN(r, S)$, is a set of $k$ points from $S$ such that $\forall p \in kNN(r, S)$, $\forall q \in \{S - kNN(r, S)\}, dist(p, r) < dist(q, r)$.*

**Definition 2.** ***AkNN:*** *Given two datasets $R, S$ and an integer $k$, the all $k$ nearest neighbors of $R$ from $S$, named $AkNN(R, S)$, is a set of pairs $(r, s)$ such that $AkNN(R, S) = \{(r, s) : r \in R, s \in kNN(r, S)\}$.*

**Definition 3.** ***AkNN Classification:*** *Given two datasets $R, S$ and a set of classes $C_S$ where points of $S$ belong, the classification process produces a set of pairs $(r, c_r)$, denoted as $AkNNC(R, S)$, such that $AkNNC(R, S) = \{(r, c_r) : r \in R, c_r \in C_S\}$ where $c_r$ is the class where the majority of $kNN(r, S)$ belong $\forall r \in R$.*

We explain Definition 3 using an illustrative example, as shown in Fig. 1. We assume that $S = \{a, b, c, d, e, f, g, h, i, j, k\}, R = \{l, m, n\}, C_S = \{A, B\}$ and

$k = 3$. We draw the boundary circle (see below in Section 3.2) that covers at least $k$ points and construct $kNN(r, S), \forall r \in R$. Next, we determine the dominant class $c_r$ in each $kNN(r, S), \forall r \in R$ and build the final $AkNNC(R, S)$ set.

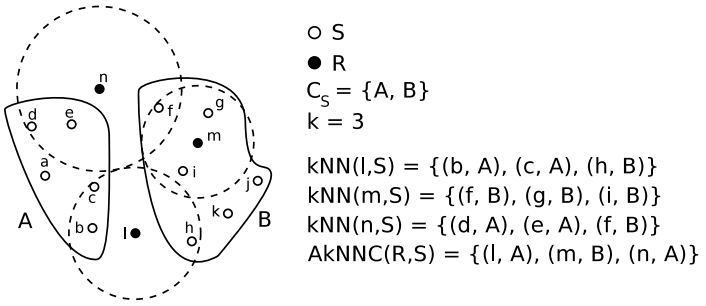



**Fig. 1.** AkNNC(R,S) explanation

### 3.2 Classification Using Space Decomposition

Consider a training dataset $T$, an input dataset $I$ and a set of classes $C_T$ where points of $T$ belong. First of all, we define as *target space* the space enclosing the points of $I$ and $T$. The partitions that are defined when we decompose the target space for 1-dimensional objects are called *intervals*. Respectively, we call *cells* and *cubes* the partitions in case of 2 and 3-dimensional objects and hypercubes for $d > 3$. For a new $1D$ point $p$, we define as *boundary interval* the minimum interval centred at $p$ that covers at least $k$-NN elements. Respectively, we define the *boundary circle* and *boundary sphere* for $2D$ and $3D$ points and the *boundary hypersphere* for $d > 3$. The notion of hypercube and hypersphere are analyzed further in Section 5. When the boundary ICSH centred in an ICCH $icch_1$, intersects the bounds of an other $icch_2$ we say an *overlap* occurs on $icch_2$. Finally, for a point $i \in I$, we define as *updates* of $kNN(i, T)$ the existence of many different instances of $kNN(i, T)$ that need to be unified to a final set.

We place the objects of $T$ on the target space according to their coordinates. The main idea of equal-sized space decomposition is to partition the target space into $n^d$ equal sized ICCHs where $n$ and the size of each ICCH are user defined. Each ICCH contains a number of points of $T$. Moreover, we construct a new layer over the target space according to $C_T$ and $\forall t \in T, c_t \in C_T$. In order to estimate $AkNNC(I, T)$, we investigate $\forall i \in I$ for $k$-nearest neighbors only in a few ICCHs, thus bounding the number of computations that need to be performed for each $i$.

### 3.3 MapReduce Model

Here, we briefly describe the MapReduce model [9]. The data processing in MapReduce is based on input data partitioning; the partitioned data is exe-

cuted by a number of tasks executed in many distributed nodes. There exist two major task categories called *Map* and *Reduce* respectively. Given input data, a *Map* function processes the data and outputs key-value pairs. Based on the Shuffle process, key-value pairs are grouped and then each group is sent to the corresponding Reduce task. A user can define his own Map and Reduce functions depending on the purpose of his application. The input and output formats of these functions are simplified as key-value pairs. Using this generic interface, the user can focus on his own problem and does not have to care how the program is executed over the distributed nodes. The architecture of MapReduce model is depicted in Fig. 2.

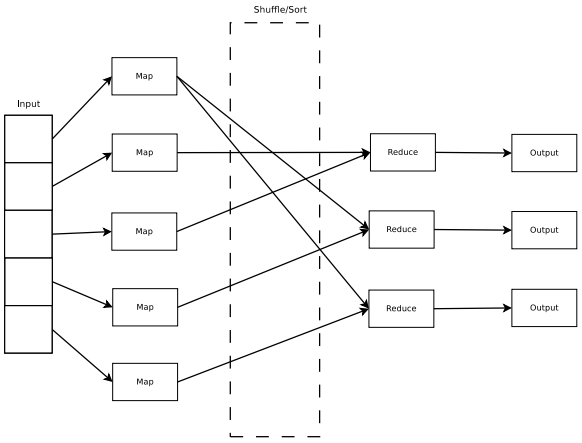



**Fig. 2.** Architecture of MapReduce model

### 3.4 Previous Work

A very preliminary study of naive A$k$NN solutions is presented in [28] and uses a simple cell decomposition technique to process A$k$NN queries on two different datasets, i.e. $I$ and $T$. The objects consisting both datasets are 2-dimensional points having only one attribute, the coordinate vector and the target space comprises of $2^n \times 2^n$ equal-sized cells.

The elements of both datasets are placed on the target space according to their coordinate vector and a cell decomposition is applied. For a point $i \in I$ it is expected that its $kNN(i, T)$ will be located in a close range area defined by nearby cells. At first, we look for candidate $k$-NN points inside the cell that $i$ belongs in the first place, name it $cl$. If we find at least $k$ elements we draw the boundary circle. There is a chance the boundary circle centred at $cl$ overlaps some neighboring cells. In this case, we need to investigate for possible $k$-NN objects inside these overlapped cells in order to create the final $k$-NN list. If no

overlap occurs, the $k$-NN list of $i$ is complete. Next, we present an example to provide a better perception of the algorithm.

Figure 3 illustrates an example of the A$k$NN process of a point in dataset $I$ using a query for $k = 3$. Initially, the point looks for $k$-NN objects inside cell 2. Since there exist at least 3 points of dataset $T$ in cell 2 the boundary circle can be drawn. The boundary circle overlaps cells 1,3 and 4, so we need to investigate for additional $k$-NN objects inside them. The algorithm outputs an instance of the $k$-NN list for every overlapped cell. These instances need to be unified into a $k$-NN list containing the final points $(x, y, z)$.

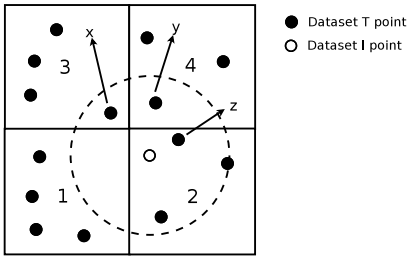



**Fig. 3.** $k$NN process using cell decomposition ($k = 3$)

This approach, as described above, fails to draw the boundary circle if $cl$ contains less than $k$ points. The solution to the problem is simple. At first, we check the number of points that fall into every cell. If we find a cell with less than $k$ points we merge it with the neighboring cells to assure that it will contain the required number of objects. The way the merging step is performed relies on the principles of hierarchical space decomposition used in quad-trees [20]. Note that this is the reason why the space decomposition involves $2^n \times 2^n$ cells. This imposes two more steps that need to be done before we begin calculating $kNN(i, T)$. In the beginning, a counting phase needs to be performed followed by a merging step in order to overcome the issue mentioned above. This preprocessing phase induces additional cost to the total computation and, as shown in the experiments, the merging step can lead to a bad algorithmic behavior.

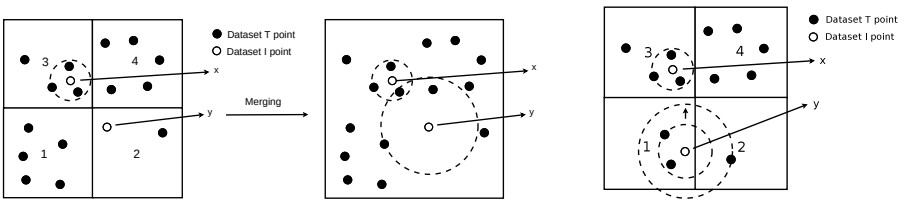
### 3.5 Technical Contributions

In this subsection, we extend the previous method for more dimensions and adapt it to the needs of the classification problem. Moreover, we analyze some drawbacks of the method studied in [28] and propose a mechanism to make the algorithm more efficient.

Firstly, we have a training dataset $T$, an input dataset $I$ and a set of classes $C_T$ where points of $T$ belong. The only difference now is that the points in the

training dataset have one more attribute, the class they belong. In order to compute $AkNNC(I,T)$, a classification step is executed after the construction of the $k$-NN lists. The class of every new object is chosen based on the class membership of its $k$-nearest neighbors. Furthermore, now the space is decomposed in $2^{dn}$ ICCHs since we consider a $d$-dimensional metric space $D$.

As mentioned before, the simple solution presented in [28] has one major drawback which is the merging step. Figure 4(a) depicts a situation where the merging step of the original method can significantly increase the total cost of the algorithm. Consider two points $x$ and $y$ entering cells 3 and 2 respectively and $k = 3$. We can draw point's $x$ boundary circle since cell 3 includes at least $k$ elements. On the contrary, we cannot draw the boundary circle of point $y$, so we need to unify cells 1 through 4 into one bigger cell. Now point $y$ can draw its boundary circle but we overload point's $x$ $k$-NN list construction with redundant computations. In the first place, the $k$-NN list of point $x$ would only need 4 distance calculations to be formed. After the merging step we need to perform 15, namely almost 4 times more than before and this would happen for all points that would join cells 1,3 and 4 in the first place.




(a) Merging issue (b) Increase range

**Fig. 4.** Issue of the merging step before the kNN process and way to avoid it ($k = 3$)

In order to avoid a scenario like above, we introduce a mechanism where only points that cannot find at least $k$-nearest neighbors in the ICCH in the first place proceed to further actions. Let a point $p$ joining an ICCH $icch$ that encloses $l < k$ neighbors. Instead of performing a merging step, we draw the boundary ICSH based on these $l$ neighbors. Then, we check if the boundary ICSH overlaps any neighboring ICCHs. In case it does, we investigate if the boundary ICSH covers at least $k$ elements in total. In case it does, then we are able to build the final $k$-NN list of the point by unifying the individual $k$-NN lists that are derived for every overlapped ICCH. In case the boundary ICSH does not cover at least $k$ objects in total or does not overlap any ICCHs, then we gradually increase its search range (by a fraction of the size of the ICCH each time) until the prerequisites are fulfilled.

Figure 4(b) explains this issue. Consider two points $x$ and $y$ entering cells 3 and 1 respectively and $k = 3$. We observe that cell 3 contains 4 neighbors and point $x$ can draw its boundary circle that covers $k$-NN elements. However, the boundary circle centred at point $y$ does not cover $k$-NN elements in the first

place. Consequently, we gradually increase its search range until the boundary circle encloses at least $k$-NN points. Note that eliminating the merging step, we also relax the condition of decomposing the target space into $2^{dn}$ equal-sized splits and generalize it to $n^d$ equal-sized splits.

Summing up, our solution can be implemented as a series of MapReduce jobs as shown below. These MapReduce jobs will be analyzed in detail in Section 4:

1. **Distribution Information**. Count the number of points of $T$ that fall into each ICCH.
2. **Primitive Computation Phase**. Calculate possible $k$-NN points $\forall i \in I$ from $T$ in the same ICCH.
3. **Update Lists**. Draw the boundary ICSH $\forall i \in I$ and increase it, if needed, until it covers at least $k$-NN points of $T$. Check for overlaps of neighboring ICCHs and derive updates of $k$-NN lists.
4. **Unify Lists**. Unify the updates of every $k$-NN list into one final $k$-NN list $\forall i \in I$.
5. **Classification**. Classify all points of $I$.

In Fig. 5, we illustrate the working flow of the A$k$NN classification process. Note, that the first MapReduce job acts as a preprocessing step and its results are provided as additional input in MapReduce Job 3 (to determine how much we need to increase the boundary ICSH) and that the preprocessing step is executed only once for $T$.

In order to fully comprehend the working flow of the A$k$NN classification process, a brief interpretation of Fig. 5 follows. In the preprocessing step, we count the number of points that fall in every ICCH (e.g. ICCH 2 contains 13 points). Now, consider points $A, C$ which belong to ICCH $2, 3$ respectively. In the second MapReduce job, we derive an initial $k$-NN list for points $A, C$ based on the objects contained in the same ICCH. In MapReduce job 3, observe that the boundary ICSH of $A$ overlaps ICCH $3, 4$ and a new $k$-NN list instance is produced for each of them. The flag in each record is *false*, which indicates the need of extra computations to build the $k$-NN list. On the other hand, the flag of $C$ equals to *true* and its $k$-NN list does not need any amendment. In MapReduce job 4, the two $k$-NN lists of $A$ are unified to a final one and lastly, in the fifth MapReduce job we classify $A, C$ based on the majority of class membership of their neighbors.

## 4 A$k$NN Classification with MapReduce

In this section, we present a detailed description of the classification process as implemented in the Hadoop framework. The whole process consists of five MapReduce jobs which are divided into three phases. Phase one estimates the distribution of $T$ over the target space. Phase two determines $kNN(i, T), \forall i \in I$ and phase three estimates $AkNNC(I, T)$. The records in $T$ have the format <point_id, coordinate_vector, class> and in $I$ have the format <point_id, coordinate_vector>. Furthermore, parameters $n$ and $k$ are defined by the user. In the

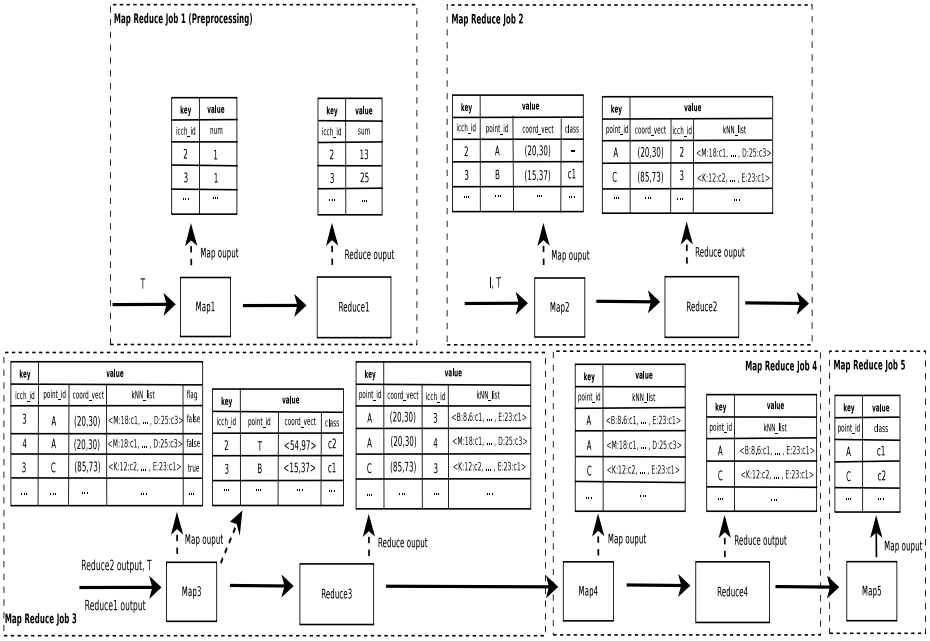



**Fig. 5.** Overview of the A$k$NN classification process

following subsections, we describe each MapReduce job separately and analyze the Map and Reduce functions that take place in each one of them. For each MapReduce job, we also quote pseudo-code, in order to provide a better comprehension of the Map and Reduce functions, and proceed to time and space complexity analysis.

### 4.1 Getting Distribution Information of Training Dataset

This MapReduce job is a preprocessing step required by subsequent MapReduce jobs that receive its output as additional data. In this step, we decompose the entire target space and count the number of points of $T$ that fall in each ICCH. Below, we sum up the Map and Reduce functions of this MapReduce process.

The *Map* function takes as input records with the training dataset format. Afterwards, it estimates the ICCH id for each point based on its coordinates and outputs a key-value pair where the key is ICCH id and the value is number 1. The *Reduce* function receives the key-value pairs from the Map function and for each ICCH id it outputs the number of points of $T$ that belong to it.

Each Map task needs $O\left(S_T/M\right)$ time to run. Each Reduce task needs $O\left(n^d/R\right)$ time to run as the total number of ICCHs is $n^d$. So, the size of the output will be $O\left(n^d \cdot c_{\text{si}}\right)$, where $c_{\text{si}}$ is the size of sum and icch_id for an output record.

**MapReduce Job 1**

```
1: function MAP(k1, v1)
2:     coord = getCoord(v1); icch_id = getId(coord);
3:     output(icch_id, 1);
4: end function

5: function REDUCE(k2, v2)
6:     sum = 0;
7:     for all v ∈ v2 do
8:         sum = sum + getSum(v);
9:     end for
10:    output(k2, sum);
11: end function
```

### 4.2 Estimating Primitive Phase Neighbors of A*k*NN Query

In this stage, we concentrate all training ($L_T$) and input ($L_I$) records for each ICCH and compute possible $k$-NN points for each item in $L_I$ from $L_T$ inside the ICCH. Below, we condense the Map and Reduce functions. We use two Map functions, one for each dataset, as seen in MapReduce Job 2 pseudo-code.

For each point $t \in T$, *Map1* outputs a new key-value pair in which the ICCH id where $t$ belongs is the key and the value consists of the id, coordinate vector and class of $t$. Similarly, for each point $i \in I$, *Map2* outputs a new key-value pair in which the ICCH id where $i$ belongs is the key and the value consists of the id and coordinate vector of $i$. The *Reduce* function receives a set of records from both Map functions with the same ICCH ids and separates points of $T$ from points of $I$ into two lists, $L_T$ and $L_I$ respectively. Then, the Reduce function calculates the distance for each point in $L_I$ from $L_T$. Subsequently, it estimates the $k$-NN points and forms a list $L$ with the format $< p_1, d_1, c_1 : \ldots : p_k, d_k, c_k >$, where $p_i$ is the $i$-th NN point, $d_i$ is its distance and $c_i$ is its class. Finally, for each $p \in L_I$, *Reduce* outputs a new key-value pair in which the key is the id of $p$ and the values comprises of the coordinate vector, ICCH id and list $L$ of $p$.

Each Map1 task needs $O(S_T/M)$ time and each Map2 task needs $O(S_I/M)$ time to run. For a Reduce task, suppose $u_i$ and $t_i$ the number of input and training points that are enclosed in an ICCH in the $i$-th execution of a Reduce function and $1 \leq i \leq n^d/R$. The Reduce task needs $O(\sum_i u_i \cdot t_i)$. Let $L_s$ to be the size of $k$-NN list and icch_id $\forall i \in I$. The output size is $O(S_I \cdot L_s) = O(S_I)$.

### 4.3 Checking for Overlaps and Updating *k*-NN Lists

In this step, at first we gradually increase the boundary ICSH (how much depends on information from the first MapReduce job), where necessary, until it includes at least $k$ points. Then, we check for overlaps between the ICSH and the neighboring ICCHs and derive updates of the $k$-NN lists. The Map and Reduce functions are outlined in MapReduce Job 3 pseudo-code. Again, we have two

**MapReduce Job 2**

```
1: function MAP1(k1, v1)
2:     coord = getCoord(v1); p_id = getPointId(v1);
3:     class = getClass(v1); icch_id = getId(coord);
4:     output(icch_id, < p_id, coord, class >);
5: end function

6: function MAP2(k1, v1)
7:     coord = getCoord(v1);
8:     p_id = getPointId(v1);
9:     icch_id = getId(coord);
10:    output(icch_id, < p_id, coord >);
11: end function

12: function REDUCE(k2, v2)
13:    L_T = getTrainingPoints(v2);
14:    L_I = getInputPoints(v2);
15:    for all p ∈ L_I do
16:        L = List{};
17:        for all t ∈ L_T do
18:            L.add(newRecord(t, dist(p, t), t.class));
19:        end for
20:        output(p.id, < p.coord, k2, getKNN(L) >);
21:    end for
22: end function
```

Map functions but the pseudo-code of Map1 function is omitted since it is the same with the respective function from MapReduce Job 2.

For each point $i \in I$, function *Map2* computes the overlaps between the ICSH and the neighboring ICCHs. If no overlap occurs, it does not need to perform any additional steps. It outputs a key-value pair in which ICCH id is the key and the value consists of id, coordinate vector and list $L$ of $i$ and a flag *true* which implies that no further process is required. Otherwise, for every overlapped ICCH it outputs a new record where ICCH id$'$ (id of an overlapped ICCH) is the key and the value consists of id, coordinate vector and list $L$ of $i$ and a flag *false*. The flag indicates we need to search for possible $k$-NN objects inside the overlapped ICCHs. The *Reduce* function receives a set of points with the same ICCH ids and separates the points of $T$ from points of $I$ into two lists, $L_T$ and $L_I$ respectively. After that, the Reduce function performs extra distance calculations using the points in $L_T$ and updates $k$-NN lists for the records in $L_I$. Finally, for each $p \in L_I$ it generates a record in which the key is the id of $p$ and the values comprises of the coordinate vector, ICCH id and list $L$ of $p$.

As before, each Map1 task needs $O\left(S_T/M\right)$ time to run. Consider an unclassified point $p$ initially belonging to an ICCH *icch*. Let $r$ be the number of times we increase the search range for $p$ and *icchov* the number of ICCHs that may be overlapped for $p$. For each Map2 task the $i$-th execution of the Map function

**MapReduce Job 3**

```
1: function MAP2(k1, v1)
2:     c = getCoord(v1); p_id = getPointId(v1);
3:     kNN = getKNNList(v1); r = getRadius(kNN);
4:     while kNN.size() < k do
5:         increase(r); kNN.addAll(getNeighbors(r));
6:     end while
7:     oICCHs = getOverlappedICCHs(r);
8:     if oICCHs.size() > 0 then
9:         for all icch ∈ oICCHs do
10:            output(icch, < p_id, c, kNN, false >);
11:        end for
12:    else
13:        output(getId(c), < p_id, c, kNN, true >);
14:    end if
15: end function

16: function REDUCE(k2, v2)
17:    L_T = getTrainingPoints(v2); L_I = getInputPoints(v2);
18:    for all p ∈ I do
19:        if p.flag == true then
20:            output(p.id, < p.coord, key, p.kNN >);
21:        else
22:            L = List{};
23:            for all t ∈ T do
24:                L.add(newRecord(t, dist(p, t), t.class));
25:            end for
26:            L_f = finalKNN(L, p.kNN);
27:            output(p.id, < p.coord, key, L_f >);
28:        end if
29:    end for
30: end function
```

performs $icchov_i + r_i$ steps, where $1 \leq i \leq S_I/M$. So, each Map2 task runs in $O\left(\sum_i (icchov_i + r_i)\right)$ time. For a Reduce task, suppose $u_i$ and $t_i$ the number of points of $I$ and $T$ respectively that are enclosed in an ICCH in the $i$-th execution of a Reduce function and $1 \leq i \leq n^d/R$. The Reduce task needs $O\left(\sum_i u_i \cdot t_i\right)$. The size of updated records is a fraction of $S_I$. So, the size of the output is also $O\left(S_I\right)$.

### 4.4 Unifying Multiple *k*-NN Lists

The previous step it is possible to yield multiple updates of a point's $k$-NN list. This MapReduce job tackles this problem and unifies possible multiple lists into one final $k$-NN list for each point $i \in I$. The Map and Reduce functions are summarized at MapReduce Job 4 pseudo-code below.

**MapReduce Job 4**

```
1: function MAP(k1, v1)
2:     output(getPointId(v1), getKKN(v1));
3: end function

4: function REDUCE(k2, v2)
5:     L = List{};
6:     for all v ∈ v2 do
7:         L.add(v);
8:     end for
9:     output(k2, unifyLists(L));
10: end function
```

The *Map* function receives the records of the previous step and extracts the $k$-NN list for each point. For each point $i \in I$, it outputs a key-value pair in which the key is the id of $i$ and the value is the list $L$. The *Reduce* function receives as input key-value pairs with the same key and computes $kNN(i, T), \forall i \in I$. The key of an output record is again the id of $i$ and the value consists of $kNN(i, T)$.

Each Map task runs in $O(S_I/M)$. For each Reduce task, assume $updates_i$ the number of updates for the $k$-NN list of an unclassified point in the $i$-th execution of a Reduce function, where $1 \leq i \leq |N_I|/R$ and $|N_I|$ the number of points in input dataset. Then, each Reduce task needs $O(\sum_i updates_i)$ to run. Let, $I_{id}$ the size of ids of all points in $I$ and $L_{final}$ is the size of the final $k$-NN list $\forall i \in I$. The size of $L_{final}$ is constant and $I_{id}$ is $O(S_I)$. Consequently, the size of the output is $O(S_I)$.

### 4.5 Classifying Points

This is the final job of the whole classification process. It is a Map-only job that classifies the input points based on the class membership of their $k$-NN points. The Map function receives as input records from the previous job and outputs $AkNNC(I, T)$. More precisely, each record handled by the *Map* function is a point together with a list of class occurrences of its $k$-NN neighbors. The function parses iteratively the list and reports the class with the highest cardinality. The key of an output record is the id of the point given as input to the Map function, while the value is the class the point is assigned. Each Map task runs in $O(S_I/M)$ time and output size is $O(S_I)$.

## 5 Extension for $d > 3$

Here we provide the extension of our method for $d > 3$. In geometry, a hypercube [7, 8] is a $n$-dimensional analogue of a square ($n = 2$) and a cube ($n = 3$) and is also called a n-cube (i.e. 0-cube is a hypercube of dimension zero and represents a point). It is a closed, compact and convex figure that consists of groups of

**MapReduce Job 5**

```
1: function MAP(k1, v1)
2:     H = HashMap < Class, Occurences > {};
3:     H = findClassOccur(v1);
4:     max = 0; maxClass = null;
5:     for all entry ∈ H do
6:         if entry.occur > max then
7:             max = entry.occur;
8:             maxClass = entry.class;
9:         end if
10:    end for
11:    output(getPointId(v1), maxClass);
12: end function
```

opposite parallel line segments aligned in each of the space's dimensions, perpendicular to each other and of the same length.

Respectively, an $n$-sphere [7, 8] is a generalization of the surface of an ordinary sphere to a $n$-dimensional space. Spheres of dimension $n > 2$ are called hyperspheres. For any natural number $n$, an $n$-sphere of radius $r$ is defined as a set of points in $(n + 1)$-dimensional Euclidean space which are at distance $r$ from a central point and $r$ may be any positive real number. So, the $n$-sphere centred at the origin is defined by:

$$S^n = \{x \in \Re^{n+1} :\| x \|= r\}$$

Figure 6 displays how to create a hypercube for $d = 4$ (4-cube) from a cube for $d = 3$. Regarding our solution for $d > 3$, the target space now is decomposed into equal-sized $d$-dimensional hypercubes and in the first place we investigate for $k$-NN points in each hypercube. Next, we draw the boundary hypersphere and increase it, if needed, until it bounds at least $k$ neighbors. Afterwards, we inspect for any overlaps between the boundary hypersphere and neighboring hypercubes. Finally, we build the final $k$-NN list for each unclassified point and categorize it according to class majority of its $k$-NN list.

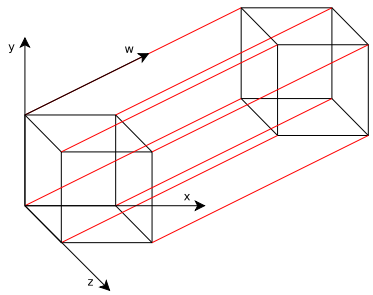



**Fig. 6.** Creating a 4-cube from a 3-cube

## 6 Experimental Evaluation

In this section, we conduct a series of experiments to evaluate the performance of our method under many different perspectives. More precisely, we take into consideration the value of $k$, granularity of space decomposition, dimensionality and data distribution.

Our cluster includes 32 computing nodes (VMs), each one of which has four 2.1 GHz CPU processors, 4 GB of memory, 40 GB hard disk and the nodes are connected by 1 gigabit Ethernet. On each node, we install Ubuntu 12.04 operating system, Java 1.7.0_40 with a 64-bit Server VM, and Hadoop 1.0.4. To adapt the Hadoop environment to our application, we apply the following changes to the default Hadoop configurations: the replication factor is set to 1; the maximum number of Map and Reduce tasks in each node is set to 3, the DFS chunk size is 256 MB and the size of virtual memory for each Map and Reduce task is set to 512 MB.

We evaluate the following approaches in the experiments:

- kdANN is the solution proposed in [28] along with the extension (which invented and implemented by us) for more dimensions, as described in Section 3, in order to be able to compare it with our solution.
- kdANN+ is our solution for $d$-dimensional points without the merging step as described in Section 3.

We evaluate our solution using both real[2] and synthetic datasets. We create $1D$ and $2D$ datasets from the real dataset keeping the $x$ and the $(x, y)$ coordinates respectively. In addition, by using statistics of underlying real dataset, we add one more dimension $z$ in order to construct a 4-dimensional dataset. We process the datasets to fit into our solution (i.e. normalization) and we end up with $1D$, $2D$, $3D$ and $4D$ datasets that consist of approximately 19,000,000 points and follow a power law like distribution. From each dataset, we extract a fraction of points (10%) that are used as a training dataset. Respectively, we create 1, 2, 3 and 4-dimensional datasets with uniformly distributed points, each dataset has 19,000,000 points and the training datasets contain 1,900,000 points. For each point in a training dataset we assign a class based on its coordinate vector. The file sizes of datasets are:

1. Real Dataset
   (a) $1D$: Input set size is 309.5 MB and training set size is 35 MB
   (b) $2D$: Input set size is 403.5 MB and training set size is 44.2 MB
   (c) $3D$: Input set size is 523.7 MB and training set size is 56.2 MB
   (d) $4D$: Input set size is 648.6 MB and training set size is 67.4 MB

[2] The real dataset is part of the Canadian Planetary Emulation Terrain 3D Mapping Dataset, which is a collection of 3-dimensional laser scans gathered at two unique planetary analogue rover test facilities in Canada. The dataset provides the coordinates $(x, y, z)$ for each laser scan in meters. `http://asrl.utias.utoronto.ca/datasets/3dmap/`

2. Synthetic Dataset
   (a) $1D$: Input set size is 300.7 MB and training set size is 33.9 MB
   (b) $2D$: Input set size is 359.2 MB and training set size is 39.8 MB
   (c) $3D$: Input set size is 478.5 MB and training set size is 51.7 MB
   (d) $4D$: Input set size is 583.4 MB and training set size is 60.9 MB

We run experiments for data up to four dimensions due to the curse of dimensionality. As shown in the experiments, for $d > 2$ the total execution cost rises exponentially and for $d > 4$ overcomes the computational power of our cluster infrastructure. We can dodge such limitations by incorporating in our system dimensionality reduction techniques, such as Principal Component Analysis (PCA) or Singular Value Decomposition [18], or elasticity mechanisms [23]. We leave this kind of extension for future work, as stated in Section 8, since it is beyond the scope of this paper.

### 6.1 Tuning parameter $n$

One major aspect in the performance of the algorithm is the tuning of granularity parameter $n$. In this experiment, we explain how to select a value of $n$ in order to succeed in achieving the shortest execution time. Each time the target space is decomposed into $2^{dn}$ equal partitions in order for kdANN to be able to perform the merging step, as described in Section 3.

In the case of power law distributions, we choose higher values of $n$ compared to uniform distributions. The intuition behind this idea, is that we want to discretize the target space into splits that contain as few points as possible in order to avoid an overload of the primitive computation phase. On the other hand, as $n$ increases, the number of update steps also increases. This can overwhelm the A$k$NN process if the number of derived instances of the $k$-NN lists is massive. Regarding uniform distributions, we wish to create larger partitions, but again not too big, in order to avoid executing many update steps. Each time, the selection of $n$ depends on the infrastructure of the cluster.

In Fig. 7, we depict how execution time varies as we alter value $n$ in case of 2-dimensional real dataset for $k = 5$. In case of kdANN+, we notice that as parameter $n$ grows the execution time drops and achieves its lowest value for $n = 9$ and slightly increases for $n = 10$. In contrary, the execution time for kdANN increases until $n = 9$ and drops significantly for $n = 10$. Moreover, its lowest achieved value is almost ten times bigger than kdANN+. Considering the above, we deduce that for power law distributions kdANN+ outperforms kdANN as $n$ changes. In addition, we conclude that the merging step affects greatly the performance of kdANN and creates a wide divergence in total running time as $n$ mutates.

Figure 8, presents the results of execution time for both methods when datasets follow a uniform distribution. Again, kdANN+ performs better than kdANN. Nevertheless, now the curve of running time presents a same behavior for both methods and when $n = 7$ the minimum running time is achieved. Observing the exported results from Fig. 7 and Fig. 8, we confirm our claim that we

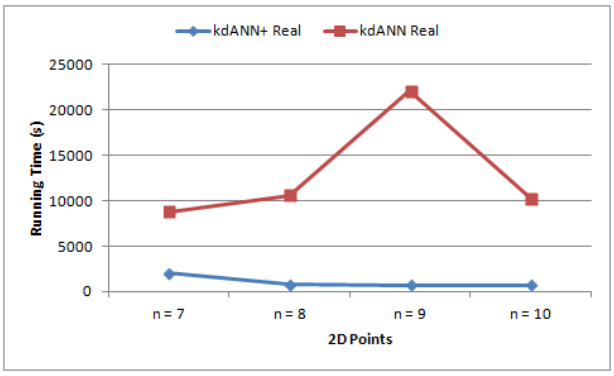



**Fig. 7.** Effect of $n$ (Real Dataset $2D$)

choose higher values of $n$ in case of power law distribution datasets, compared to uniformly distributed datasets, in order to minimize the total execution time.

We proceed to similar experimental procedures for all dimensions. The results for $1D$, $3D$ and $4D$ points follow the same trend (we omit the graphs of other dimensions to avoid pointless repetition). In the case of real datasets, we pick value $n$ that maximizes the performance of kdANN+ since kdANN presents a bad algorithmic behavior regardless of value $n$, as shown in the majority of experiments that follow.

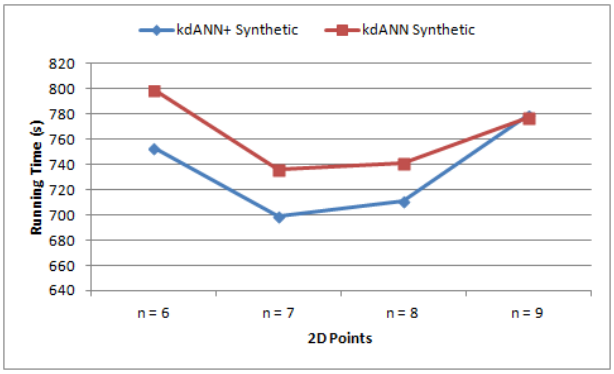



**Fig. 8.** Effect of $n$ (Synthetic Dataset $2D$)

### 6.2 Effect of $k$ and Effect of Dimensionality

In this experiment, we evaluate both methods using real and synthetic datasets and record the execution time as $k$ increases for each dimension. Finally, we

study the effect of dimensionality on the performance of kdANN and kdANN+. Based on the findings of Section 6.1, for the rest of our experiments we set the value $n$ as summarized below:

1. Real Dataset
   (a) $1D$: $n = 18$
   (b) $2D$: $n = 9$
   (c) $3D$: $n = 7$
   (d) $4D$: $n = 6$
2. Synthetic Dataset
   (a) $1D$: $n = 16$
   (b) $2D$: $n = 7$
   (c) $3D$: $n = 5$
   (d) $4D$: $n = 4$

**Effect of $k$ for Different Dimensions.** Figure 9 presents the results for kdANN and kdANN+ by varying $k$ from 5 to 20 on real and synthetic datasets. In terms of running time, kdANN+ always perform better, followed by kdANN and each method behave in the same way for both datasets, real and synthetic. As the value of $k$ grows, the size of each intermediate record becomes larger respectively. Consequently, the data processing time increments. Moreover, as the number of neighbors we need to estimate each time augments, we need to search into more intervals for possible $k$-NN points as the boundary interval grows larger.

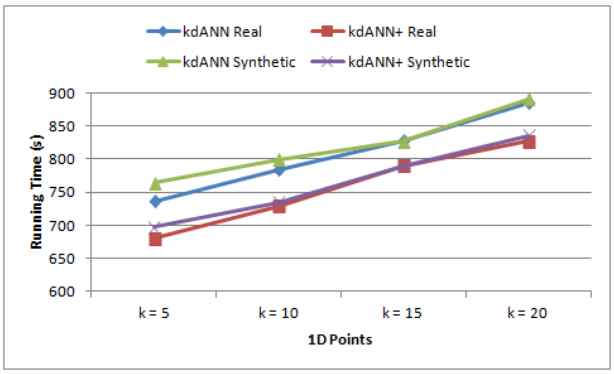



**Fig. 9.** Effect of $k$ for $d = 1$

In Fig. 10, we demonstrate the outcome of the experimental procedure for 2-dimensional points when we alter $k$ value from 5 to 20. First of all, note that we do not include the results of kdANN for the real dataset. The reason is that the method only produced results for $k = 5$ and needed more than 4

hours. Beyond this, the merging step of kdANN derived extremely sizeable cells. Consequently, during the primitive computation phase a bottleneck was created to some nodes that strangled their resources, thus preventing them to yield any results. Observing the rest of the curves, we notice that the processing times are a bit higher than the previous ones due to larger records, as we impose one more dimension. Furthermore, the search area now overlaps more partitions of the target space than in case of 1-dimensional points. Consequently, the algorithm produces more instances of the $k$-NN lists and the time requirement to merge them is bigger. Overall, in the case of power law distribution, kdANN+ behaves much better than kdANN since the last one fails to process the A$k$NN query as $k$ increases. Also, kdANN+ is faster and in case of synthetic dataset that follows a uniform distribution, especially as $k$ increases.

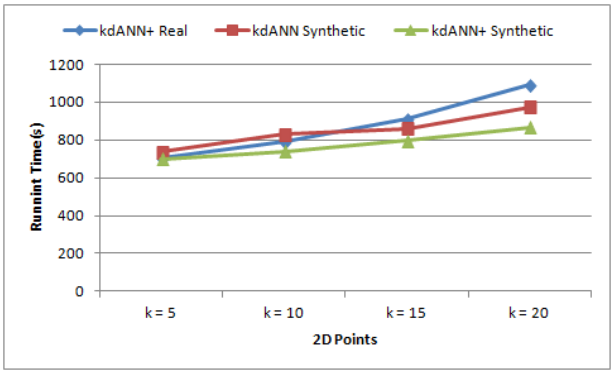



**Fig. 10.** Effect of $k$ for $d = 2$

Figure 11 displays the results generated from kdANN and kdANN+ for the 3-dimensional points when we increase $k$ value from 5 to 20. Once again, in case of kdANN we could not get any results for any value of $k$ when we provided the real dataset as input. The reasons are the same we mentioned in the previous paragraph for $d = 2$.

Table 2 is pretty illustrative in the way the merging step affects the A$k$NN process. First of all, its computational cost is far from negligible if performed in a node (in contrary with the claim of the authors as stated in [28]). Apart from this, the ratio of cubes that participate in the merging process is almost 40% and the largest merged cube consists of 32,768 and 262,144 initial cubes for $k = 5$ and $k > 5$ respectively.

In the case of kdANN+, when given the real dataset as input, it is obvious that the total computational cost is much larger compared to the one shown in Figures 9 and 10. This happens for 3 reasons: 1) we have larger records in size, 2) some cubes are quite denser compared to others (since the dataset follows a power law distribution) and we need to perform more computations for them in the primitive computation phase and 3) a significant amount of overlaps take

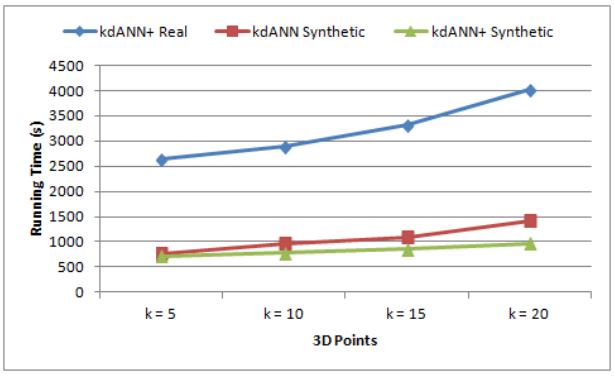



**Fig. 11.** Effect of $k$ for $d = 3$

**Table 2.** Statistics of merging step for kdANN

| | $k = 5$ | $k = 10$ | $k = 15$ | $k = 20$ |
|---|---|---|---|---|
| Time (s) | 271 | 675 | 962 | 1,528 |
| # of merged cubes | 798,032 | 859,944 | 866,808 | 870,784 |
| % of total cubes | 38% | 41% | 41.3% | 41.5% |
| Max merged cubes | 32,768 | 262,144 | 262,144 | 262,144 |

place, thus the update step of the $k$-NN lists needs more time than before. Finally, kdANN+ performs much better than kdANN, in the case of synthetic dataset, and the gap between the curves of running time tends to be bigger as $k$ increases.

Finally, Fig. 12 demonstrates the total running cost for both kdANN and kdANN+ in the case of 4-dimensional datasets. Our method kdANN+, continues to overrun kdANN when our input follows a uniform distribution and the variance between the curves is a bit bigger than the previous cases. As expected, kdANN flunks in producing results for the real dataset while our method answers the AkNN query requiring much more processing time than the 3-dimensional case. The curve has a tendency to increase exponentially (we explained the reasons in the previous paragraph) and for $k = 20$ the time taken to export the outcome of the AkNN query is almost double compared to the running time of Fig. 11.

**Effect of Dimensionality.** In this subsection, we evaluate the effect of dimensionality for both real and synthetic datasets. Figure 13 presents the running time for $k = 20$ by varying the number of dimensions from 1 to 4.

From the outcome, we observe that kdANN is more sensitive to the number of dimensions than kdANN+ when we provide a dataset with uniform distribution as input. In particular, when the number of dimensions varies from 2 to 4 the divergence between the two curves starts growing faster. In the case of power law distribution, we only include the results for kdANN+ since kdANN fails to

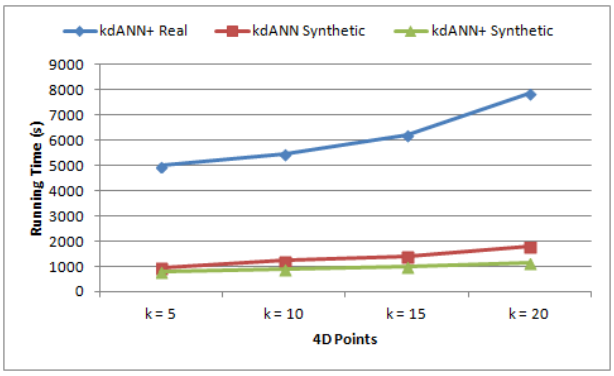



**Fig. 12.** Effect of $k$ for $d = 4$

process the A$k$NN query for dimensions 2 to 4 when $k = 20$. We notice that the execution time increases exponentially when $d > 2$. This results from the curse of dimensionality. As the number of dimensions increases, the number of distance computations as well as the number of searches in neighboring ICCHs increases exponentially. Nevertheless, kdANN+ can still process the A$k$NN query in a reasonable amount of time in contrast to kdANN.

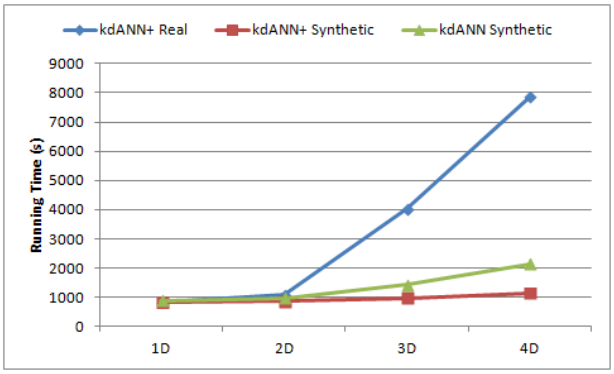
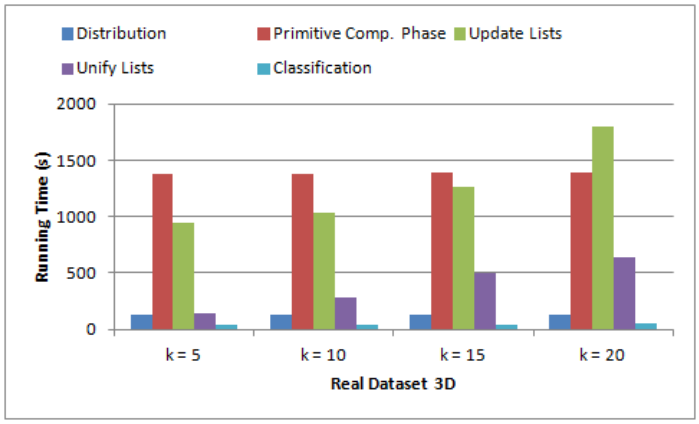


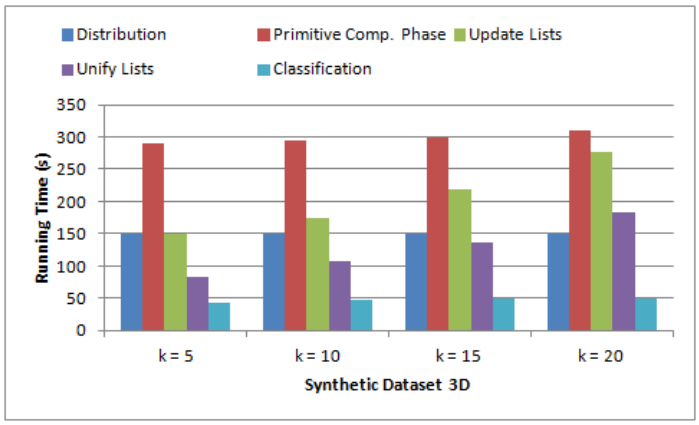

**Fig. 13.** Effect of dimensionality for $k = 20$

### 6.3 Phase Breakdown.

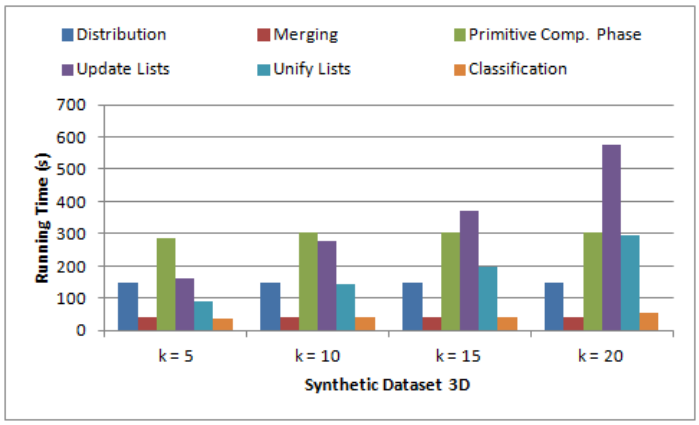
In Figures 14(a)-14(c) we present the results of running time for different stages of kdANN and kdANN+, in case of 3-dimensional datasets, as $k$ increases. We observe, that in all figures, the running time of distribution phase is the same (it runs only once since it is a preprocessing step). On the other hand, the running




(a) kdANN+




(b) kdANN+




(c) kdANN

**Fig. 14.** Phase breakdown vs $k$

time of primitive computational and classification phase slightly increase as $k$ grows. Since the $k$-NN list gets bigger, the algorithm takes more time to process the input records. Considering update and integrate phase, the running time increases notably. The bigger the value of $k$, the bigger the cardinality of derived instances of the $k$-NN lists due to larger area coverage by the boundary ICSH. Consequently, the algorithm needs more time to derive the final $k$-NN lists. The cumulative cost of these two phases is the one that mostly affects the total running time of the A$k$NN query in the majority of the experiments. Finally, the execution time of the merging phase remains the same (in case of kdANN in Fig. 14(c)). Apart from the merging phase, whose running cost may increase significantly (Table 2), the execution time for the rest phases follow the same trend as $d$ varies.

### 6.4 Power Law vs Uniform Distribution.

In this subsection, we perform a comparative analysis of the results exported by our method for datasets with different distributions and argue about the performance of methods kdANN and kdANN+ as $k$ and $d$ increments.

At first, we observe that as $k$ increases kdANN+ prevails kdANN for all dimensions and for both dataset distributions (based on results from Figs. 9 - 12). It is clear that our contribution presented in Section 3.4, speeds up the solution presented in [28]. Under the perspective of dimensionality, in case of uniform distribution the divergence between the curves is not very big. Nevertheless, the running time of kdANN+ increases linearly whilst kdANN's running time grows exponentially for $d > 2$ (see Fig. 13). On the other hand, in case of power law distribution, for $d > 1$ kdANN+ outperforms kdANN. The last one either fails to derive results in a reasonable amount of time or cannot produce any results at all (again see Fig. 13). As shown in Table 2, the merging step has major deficiencies. It can cumber with notable computational burden the total A$k$NN process and can produce quite large merged ICCHs. As a consequence, the workload is badly distributed among the nodes and some of them end up running out of resources, thus causing kdANN to fail to produce any results. Despite the superiority of kdANN+, its execution time increases exponentially when the number of dimensions varies from 2 to 4.

Overall, the experimental evaluation shows that our solution (kdANN+) scales better than kdANN for uniform distributions and dominates it for power law distributions. However, it is clear that both kdANN and kdANN+ are more sensitive to power law distributions. As a result, their performance degrades faster than the case of uniform distributions.

### 6.5 Scalability

In this experiment, we investigate the scalability of the two approaches. We utilize the $3D$ datasets, since their size is quite big, and create new chunks smaller in size that are a fraction $F$ of the original datasets, where $F \in \{0.2, 0.4, 0.6, 0.8\}$. Moreover, we set the value of $k$ to 5.

Figure 15 presents the scalability results for real and synthetic datasets. In the case of power law distribution, the results display that kdANN+ scales almost linearly as the data size increases. In contrast, kdANN fails to generate any results even for very small datasets since the merging step continues to be an inhibitor factor in kdANN's performance. In addition, we can see that kdANN+ scales better than kdANN in the case of synthetic dataset and the running time increases almost linearly as in the case of power law distribution. Regarding kdANN, the curve of execution time is steeper until $F = 0.6$ and after that it increases more smoothly.

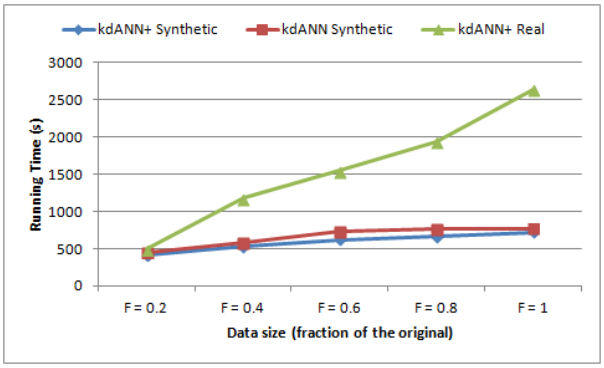



**Fig. 15.** Scalability

Table 3 shows the way the merging step affects kdANN as the data size varies. The ratio of cubes that are involved in the merging process remains high and varies from 36.6% to 39.3% and the largest merged cube comprises of 32,768 cubes of the initial space decomposition. Interestingly, the time to perform the merging step is not strictly increasing proportionally to the data size. In particular, the worst time is achieved when $F = 0.2$, then it reaches its minimum value for $F = 0.4$ and beyond this value augments again. Below, we explain why this phenomenon appears. The merging process takes into account the distribution information of dataset $T$. As the size of the input dataset decreases, respectively the size of the training dataset also mitigates. Since both datasets follow a power law distribution, the ICCHs that include training set points decrease also in number and this may result in more merging steps (i.e. $F = 0.2$).

### 6.6 Speedup

In our last experiment, we measure the effect of the number of computing nodes. We test four different cluster configurations and the cluster consist of $N \in \{11, 18, 25, 32\}$ nodes each time. We test the cluster configurations against the 3-dimensional datasets when $k = 5$.

**Table 3.** Statistics of merging step for kdANN and different data sizes

| | $F = 0.2$ | $F = 0.4$ | $F = 0.6$ | $F = 0.8$ |
|---|---|---|---|---|
| Time (s) | 598 | 223 | 279 | 300 |
| # of merged cubes | 825,264 | 767,768 | 768,256 | 802,216 |
| % of total cubes | 39.3% | 36.6% | 36.6% | 38.2% |
| Max merged cubes | 32,768 | 32,768 | 32,768 | 32,768 |

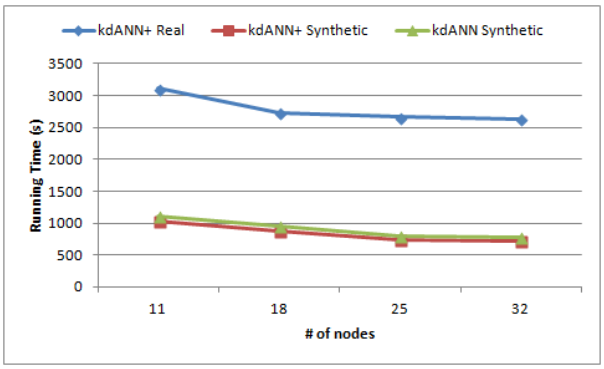



**Fig. 16.** Speedup

From Fig. 16, we observe that total running time of kdANN+, in the case of power law distribution, tends to decrease as we add more nodes to the cluster. Due to the increment of number of computing nodes, the amount of distance calculations and update steps on $k$-NN lists that undertakes each node decreases respectively. Moreover, since kdANN fails to produce any results using 3-dimensional real dataset when the cluster consists of 32 nodes, it is obvious that it will fail with less nodes too. That is the reason for the absence of kdANN's curve from Fig. 16. In the case of synthetic dataset, we observe that both kdANN and kdANN+ achieve almost the same speedup as the number of nodes increases; still kdANN+ performs betters than kdANN. We behold that in the case of real dataset the curve of running time decreases steeper as the number of nodes varies from 11 to 18 and becomes smoother beyond this point. On the other hand, in case of synthetic dataset the curves decrease smoother when the number of nodes varies from 25 to 32. The conclusion that accrues from this observation is that the increment of computing nodes has a greater effect on the running time of both approaches when the datasets follow a uniform distribution. This happens because the workload is distributed better among the nodes of the cluster.

### 6.7 Classification Performance

In this section, we present the performance results of our classification method for kdANN+, when the $3D$ real dataset is provided as input and $k = 10$. We define

a set $C_T = \{A, B, C, D, E\}$ of 5 classes over the target space, but only 3 of them $(A, B, C)$ contain points of $T$. The class where a point $t \in T$ belongs, depends on its coordinate vector. In Table 4, we measure the classification performance using four metrics for each class, *True Positive, False Negative, False Positive* and *True Negative* and give an average on the performance of each metric for all the classes. Among the classes, class $C$ has the worst accuracy but the overall results show that our classification method performs well.

**Table 4.** Classification performance of kdANN+

| | Class $A$ | Class $B$ | Class $C$ | Average |
|---|---|---|---|---|
| True Positive | 99.97% | 99.91% | 94.14% | 98% |
| False Negative | 0.03% | 0.09% | 5.86% | 2% |
| False Positive | 0.13% | 0.06% | 0.06% | 0.08% |
| True Negative | 99.87% | 99.94% | 99.94% | 99.92% |

## 7 Conclusions

In the context of this work, we presented a novel method for classifying multidimensional data using A$k$NN queries in a single batch-based process in Hadoop. To our knowledge, it is the first time a MapReduce approach for classifying multidimensional data is discussed. By exploiting equal-sized space decomposition techniques we bound the number of distance calculations we need to perform for each point to reckon its $k$-nearest neighbors. We conduct a variety of experiments to test the efficiency of our method on both, real and synthetic datasets. Through this extensive experimental evaluation we prove that our system is efficient, robust and scalable.

## 8 Future Work

In the near future, we plan to extend and improve our system in order to boost its efficiency and flexibility. At first, we want to relax the condition of decomposing the target space into equal-sized splits. We have in mind to implement a technique that will allow us to have unequal splits, containing approximately the same number of points. This is going to decrease the number of overlaps and calculations for candidate $k$-NN points. Moreover, the method will become distribution independent leading to better load balancing between the nodes.

In addition, we intend to apply a mechanism in order for the cluster to be used in a more elastic way, by adding (respectively removing) nodes as the number of dimensions increase (respectively decrease) or the data distribution becomes more (respectively less) challenging to handle.

Finally, we plan to use indexes, such as R-trees or M-trees, along with HBase, in order to prune any points that are redundant and cumber additional cost to the method.

**Acknowledgements.** This research has been co-financed by the European Union (European Social Fund ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thales. Investing in knowledge society through the European Social Fund.

## References

1. Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X., Saltz, J.: Hadoop GIS: A High Performance Spatial Data Warehousing System over Mapreduce. Proc. VLDB Endow. 6, 1009–1020 (2013)
2. Afrati, F.N., Ullman, J.D.: Optimizing Joins in a Map-Reduce Environment. In: Proceedings of the 13th International Conference on Extending Database Technology, pp. 99–110. ACM, New York, NY, USA (2010)
3. Böhm, C., Krebs, F.: The k-Nearest Neighbour Join: Turbo Charging the KDD Process. Knowl. Inf. Syst. 6, 728–749 (2004)
4. Chatzimilioudis, G., Zeinalipour-Yazti, D., Lee, W.-C., Dikaiakos, M. D.: Continuous All k-Nearest-Neighbor Querying in Smartphone Networks. In: Proceedings of the 2012 IEEE 13th International Conference on Mobile Data Management, pp. 79–88. IEEE Computer Society, Washington, DC, USA (2012)
5. Chang, J., Luo, J., Huang, J.Z., Feng, S., Fan, J.: Minimum Spanning Tree Based Classification Model for Massive Data with MapReduce Implementation. In: Proceedings of the 10th IEEE International Conference on Data Mining Workshop, pp. 129–137. IEEE Computer Society, Washington, DC, USA (2010)
6. Chen, Y., Patel, J.M.: Efficient Evaluation of All-Nearest-Neighbor Queries. In: Proceedings of the 23rd IEEE International Conference on Data Engineering, pp. 1056–1065. IEEE Computer Society, Washington, DC, USA (2007)
7. Coxeter, H.S.M.: Regular Polytopes. Dover Publications (1973)
8. Cromwell, P.R.: Polyhedra. Cambridge University Press (1999)
9. Dean, J., Ghemawat, S.: MapReduce: Simplified Data Processing on Large Clusters. In: Proceedings of the 6th Symposium on Operating Systems Design and Implementation, pp. 137–150. USENIX Association, Berkeley, CA, USA (2004)
10. Dunham, M.H.: Data Mining, Introductory and Advanced Topics. Prentice Hall, Upper Saddle River, NJ, USA (2002)
11. Eldawy, A.: SpatialHadoop: Towards Flexible and Scalable Spatial Processing Using Mapreduce. In: Proceedings of the 2014 SIGMOD PhD Symposium, pp. 46–50. ACM, New York, NY, USA (2014)
12. Emrich, T., Graf, F., Kriegel, H.-P., Schubert, M., Thoma, M.: Optimizing All-Nearest-Neighbor Queries with Trigonometric Pruning. In: Scientific and Statistical Database Management. LNCS, vol. 6187, pp. 501–518. Springer-Verlag, Berlin, Heidelberg (2010)um, pp. 46–50. ACM, New York, NY, USA (2014)
13. Gkoulalas-Divanis, A., Verykios, V.S., Bozanis, P.: A Network Aware Privacy Model for Online Requests in Trajectory Data. Data Knowl. Eng. 68, 431–452 (2009)
14. He, Q., Zhuang, F., Li, J., Shi, Z.: Parallel implementation of classification algorithms based on MapReduce. In: Proceedings of the 5th International Conference on Rough Set and Knowledge Technology, pp. 655–662. Springer-Verlag, Berlin, Heidelberg (2010)

15. Ioup, E., Shaw, K., Sample, J., Abdelguerfi, M.: Efficient AKNN spatial network queries using the M-Tree. In: Proceedings of the 15th annual ACM International Symposium on Advances in Geographic Information Systems, pp. 46:1–46:4. ACM, New York, NY, USA (2007)
16. Lee, K., Ganti, R.K., Srivatsa, M., Liu, L.: Efficient spatial query processing for big data. In: Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 469–472. ACM, New York, NY, USA (2014)
17. Lu, W., Shen, Y., Chen, S., Ooi, B.C.: Efficient Processing of k Nearest Neighbor Joins using MapReduce. Proc. VLDB Endow. 5, 1016–1027 (2012)
18. Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press (2011)
19. Roussopoulos, N., Kelley, S., Vincent, F.: Nearest Neighbor Queries. In: Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data, pp. 71–79. ACM, New York, NY, USA (1995)
20. Samet, H.: The QuadTree and Related Hierarchical Data Structures. ACM Comput. Surv. 16, 187–260 (1984)
21. Stupar, A., Michel, S., Schenkel, R.: RankReduce - Processing K-Nearest Neighbor Queries on Top of MapReduce. In: Proceedings of the 8th Workshop on Large-Scale Distributed Systems for Information Retrieval, pp. 13–18. (2010)
22. The apache software foundation: Hadoop homepage, `http://hadoop.apache.org/`
23. Tsoumakos, D., Konstantinou, I., Boumpouka, C., Sioutas, S., Koziris, N.: Automated, Elastic Resource Provisioning for NoSQL Clusters Using TIRAMOLA. In: Proceedings of the 13th IEEE/ACM International Symposium on Cluster, Cloud, and Grid Computing, pp. 34–41. (2013)
24. Vernica, R., Carey, M.J., Li, C.: Efficient Parallel Set-Similarity Joins Using MapReduce. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 495–506. ACM, New York, NY, USA (2010)
25. White, T.: Hadoop: The Definitive Guide, 3rd Edition. O'Reilly Media / Yahoo Press (2012)
26. Xia, C., Lu, H., Chin, B., Hu, O.J.: Gorder: An efficient method for knn join processing. In: VLDB, pp. 756–767. VLDB Endowment (2004)
27. Yao, B., Li, F., Kumar, P.: K Nearest Neighbor Queries and KNN-Joins in Large Relational Databases (Almost) for Free. In: Proceedings of the 26th International Conference on Data Engineering, pp. 4–15. IEEE Computer Society, Washington, DC, USA (2010)
28. Yokoyama, T., Ishikawa, Y., Suzuki, Y.: Processing All k-Nearest Neighbor Queries in Hadoop. In: Proceedings of the 13th International Conference on Web-Age Information Management. LNCS, vol. 7418, pp. 346–351 (2012)
29. Yu, C., Cui, B., Wang, S., Su, J.: Efficient index-based KNN join processing for high-dimensional data. Information & Software Technology 49, 332–344 (2007)
30. Zhang, C., Li, F., Jestes, J.: Efficient Parallel kNN Joins for Large Data in MapReduce. In: Proceedings of the 15th International Conference on Extending Database Technology, pp. 38–49. ACM, New York, NY, USA (2012)
31. Zhang, J., Mamoulis, N., Papadias, D., Tao, Y.: All-Nearest-Neighbors Queries in Spatial Databases. In: Proceedings of the 16th International Conference on Scientific and Statistical Database Management, pp. 297–306. IEEE Computer Society, Washington, DC, USA (2004)