Understanding the SNN Input Parameters and How They Affect the Clustering Results

Guilherme Moreira, ALGORITMI Research Centre, University of Minho, Guimarães, Portugal

Maribel Yasmina Santos, ALGORITMI Research Centre, University of Minho, Guimarães, Portugal

João Moura Pires, NOVA LINCS, Nova University of Lisbon, Lisbon, Portugal

João Galvão, ALGORITMI Research Centre, University of Minho, Guimarães, Portugal

ABSTRACT

Huge amounts of data are available for analysis in nowadays organizations, which are facing several challenges when trying to analyze the generated data with the aim of extracting useful information. This analytical capability needs to be enhanced with tools capable of dealing with big data sets without making the analytical process an arduous task. Clustering is usually used in the data analysis process, as this technique does not require any prior knowledge about the data. However, clustering algorithms usually require one or more input parameters that influence the clustering process and the results that can be obtained. This work analyses the relation between the three input parameters of the SNN (Shared Nearest Neighbor) clustering algorithm, providing a comprehensive understanding of the relationships that were identified between k, Eps and MinPts, the algorithm’s input parameters. Moreover, this work also proposes specific guidelines for the definition of the appropriate input parameters, optimizing the processing time, as the number of trials needed to achieve appropriate results can be substantially reduced.

Keywords: Density-Based Clustering, Input Parameters Tuning, Shared Nearest Neighbor

1. CONTEXT AND MOTIVATION

Current technological developments allow the collection of huge amounts of data that are usually stored and analyzed to support the decision-making process. Analytical tools, like data mining algorithms support the analysis of such vast amount of data. Data mining is one of the steps of the knowledge discovery process (Fayyad, Piatetsky-Shapiro, Smyth, & Uturu samy, 1996), in which clustering algorithms are common techniques used to analyze data, not requiring any prior knowledge about the data set (Jain, Murty, & Flynn, 1999). Being unsu-

DOI: 10.4018/IJDWM.2015070102
supervised data mining techniques, clustering has the advantage of identifying clusters that emerge naturally from data. However, most of the clustering algorithms require input parameters that influence the results that can be obtained. The process of tuning the input parameters is usually a trial and error process in which the user changes the input parameters until the results satisfy the analysis requirements (Bouguessa, 2011). This process can be difficult and time consuming as no strict rules exist about the definition of these parameters. Moreover, any new trial requires a new run of the algorithm so more processing time is needed. To overcome this trial and error process, this work analyses in detail the three input parameters – $k$, $Eps$ and $MinPts$ – of the SNN (Shared Nearest Neighbor) algorithm and proposes specific guidelines to identify these parameters, which are used to calibrate the clustering’s results, attending to the data available for analysis. $k$ represents the size of nearest neighbors list; $Eps$ is the density threshold; and, $MinPts$ is the minimum density that a point needs to satisfy to become a core point of a cluster (Ertoz, Steinbach, & Kumar, 2003).

The problem of input parameters identification was addressed in previous works (Birant & Kut, 2007; Ester, Kriegel, Sander, & Xu, 1996; Guha, Rastogi, & Shim, 1998; Karypis, Han, & Kumar, 1999). Birant (Birant & Kut, 2007) suggested that $k$ can be approximated using the natural logarithm of the number of objects to be clustered ($ln(n)$). Although providing a good approximation for small data sets, the value of $k$ obtained for large data set is not appropriate, as this paper will show. Ertoz (Ertoz et al., 2003) indicates that knowing the value of $k$, the $MinPts$ input parameter can be calculated as being a fraction of $k$. For $Eps$, some heuristics have been proposed mainly for the DBSCAN algorithm (Ester et al., 1996), but they cannot be applied to SNN as the $Eps$ parameter in SNN is used with a different semantic.

This paper is organized as follows. Section 2 summarizes related work on density-based clustering approaches, and the SNN algorithm, and their attempt to identify the input parameters. Section 3 analyses synthetic data sets, establishing relationships between the input parameters. Section 4 presents a comprehensive overview of the $Eps$ input parameter. Section 5 concludes with a summary of the main findings and proposals of future work.

2. RELATED WORK

Clustering is the task of identifying sets of segments or clusters that group similar objects. A cluster is a collection of data objects that have more similarities between them and are dissimilar to objects that belong to other clusters (Han & Kamber, 2001).

Density-based clustering approaches were developed based on the notion of density (Han & Kamber, 2001). These algorithms perceive clusters as dense regions of objects in a space separated by regions of relatively low density. This kind of algorithms is useful to filter out noise and for discovering clusters of arbitrary shapes (Ye & others, 2003). DBSCAN (Ester et al., 1996) and OPTICS (Ankerst, Breunig, Kriegel, & Sander, 1999) are major representatives of this class of clustering algorithms, being DBSCAN the most representative density-based clustering algorithm. Many of the available density-based algorithms were derived from DBSCAN, which was introduced by Ester (Ester et al., 1996) and was specially designed to treat spatial databases.

DBSCAN needs two input parameters: $Eps$ and $MinPts$. $Eps$ is the radius distance of a point, in other words the neighborhood of a point. $MinPts$ is the minimum number of points that the neighborhood must have to be considered a cluster. Points within the radius are considered core points and points on the border of the cluster are border points. Points that do not belong to any cluster are considered noise points (Ester et al., 1996).

The SNN algorithm is an extension of the Jarvis-Patrick (Jarvis & Patrick, 1973) and the DBSCAN (Ester et al., 1996) algorithms. The main difference and contribution from (Ertoz et al., 2003) is the similarity measure implemented.
Online Processing of End-User Data in Real-Time Data Warehousing

Security of Wireless Devices using Biological-Inspired RF Fingerprinting Technique