Pervasive and Mobile Computing 28 (2016) 69–80

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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

Improving biomedical signal search results in big data case-based reasoning environments

Jonathan Woodbridge[a], Bobak Mortazavi[a,*], Alex A.T. Bui[b], Majid Sarrafzadeh[a]

[a] Computer Science Department, UCLA, Los Angeles, CA 90095, United States
[b] Medical Imaging Informatics, UCLA, Los Angeles, CA 90095, United States

ARTICLE INFO

Article history:
Available online 27 October 2015

Keywords:
Biomedical signal search
Locality Sensitive Hashing
Time-series subsequence matching
Monte Carlo sampling
Case-based reasoning

ABSTRACT

Time series subsequence matching has importance in a variety of areas in healthcare informatics. These include case-based diagnosis and treatment as well as discovery of trends among patients. However, few medical systems employ subsequence matching due to high computational and memory complexities. This paper proposes a randomized Monte Carlo sampling method to broaden search criteria with minimal increases in computational and memory complexities over *R*-NN indexing. Information gain improves while producing result sets that approximate the theoretical result space, query results increase by several orders of magnitude, and recall is improved with no significant degradation to precision over *R*-NN matching.

1. Introduction

Medical case-based reasoning (CBR) is a well-studied area for medical diagnosis and treatments [1–3]. These systems rely on data mining and machine learning techniques to derive decisions on new cases based on a knowledge-base of previous cases. One major drawback of CBR systems is that the knowledge base must contain relevant cases to correctly make decisions on a current case [3,4]. This is especially difficult for high dimensional measured signals, such as Electrocardiogram (ECG) [5] and accelerometers which exhibit high variability due to noise, sensor displacement, misuse, and other factors that are difficult to control [6]. Systems hope to use such information to accurately classify signals and patients for an accurate medical diagnosis [7], dealing with the complexities as well as the potential issues of missing data [8]. A CBR system based on high dimensional measured signals must be extremely large to not only account for the variability of patients, but to the variability of the signal type. The memory and computational complexity of such systems can limit the information gain provided. Indeed, as medical devices produce larger quantities of data and more frequently, efficient search becomes paramount in identifying important information and searching for useful and related signals. This work will investigate not only the quality of information presented in the signal search engine developed, but also the speed with which such a system returns results, for usefulness in a case-based reasoning environment.

* Corresponding author.
E-mail addresses: jwoodbri@cs.ucla.edu (J. Woodbridge), bobakm@cs.ucla.edu (B. Mortazavi), buia@mii.ucla.edu (A.A.T. Bui), majid@cs.ucla.edu (M. Sarrafzadeh).

http://dx.doi.org/10.1016/j.pmcj.2015.09.006
1574-1192/© 2015 Elsevier B.V. All rights reserved.

High dimensional subsequence matching, or R nearest neighbor (R-NN), is the process of finding similar segments within a database of high dimensional measured signals. A match is defined as any two segments $u, v \in S$ such that $dist(u, v) \leq R$ where S is the search space of biomedical signals, *dist* is a measure of distance between two signals (such as Euclidean distance), and R is a predefined threshold. In practice, R tends to be relatively small, leading to homogeneous result sets. While results may be precise, they offer little information gain. Of course, result sets can be enlarged by increasing R. However, arbitrarily increasing R can destabilize the result set, rendering it meaningless [9].

This paper presents a randomized Monte Carlo approach for improving R-NN search results. This approach enlarges search results while ensuring precision and yielding higher relative information gain. The method is built upon two assumptions: time series databases are extremely large [10] and result sets follow a Gaussian distribution [11,12]. The proposed method consists of two steps. First, a query segment q undergoes m randomizations constructing a set Q of query segments where $|Q| = m$. Next, an R-NN search for each $u \in Q$ is performed using the l_2 norm (Euclidean distance). The Euclidean distance between q and all segments $u \in Q$ follows a Gaussian distribution with a mean μ_Q and standard deviation σ_Q determined by the randomization.

There are several R-NN methods that exist in the literature including spatial indexes [13–15] and Locality Sensitive Hashing (LSH) [16–18]. This paper utilizes LSH as the underlying hash-based nearest neighbor search algorithm, but the theoretical contributions of this paper are applicable to most R-NN methods. However, the optimizations proposed by this paper were designed predominately for an LSH scheme.

Results from this paper are shown both theoretically and experimentally. Experiments are run on both synthetic random walk and real-world publicly available datasets. The randomized approach increased the number of search results by several orders of magnitude over LSH alone while keeping similar preciseness. Experimental databases contained tens of millions of indexed sub-sequences showing both correctness and scalability. However, the proposed algorithm is highly parallelizable, potentially allowing for databases of a much larger scale. The results present important information for a case-based reasoning system.

This paper is an extension of [19] in which further investigation of case-based reasoning environments is presented, as well as scope and evaluation of the method is extended to better represent the performance improvements; results generated in this work also consider the wall-clock time taken in order to perform the tasks outlined. The rest of the paper is organized as follows: Section 2 presents the motivation and related work in signal searching; Section 3 provides the proposed method as well as its theoretical proof; and Section 3.4 describes the experimental set-up with the results and relating discussion in Section 4. Conclusions are given in Section 5.

2. Background

2.1. Biomedical signals

Many different strategies to searching efficiently through biomedical time series exist, particularly with ECG and EEG signals. Authors in [20] attempt to approach the problem from a multi-dimensional angle. However, their search uses a dynamic time warping approach to create exact matches for plantar pressure. The proposed method accounts for an extension beyond dynamic time warping and is not truly adaptable to a large database for case-based reasoning due to its time-complexity and information in its results. Many works in ECG and EEG, such as those by authors in [21] and [22] respectively focus on analyzing time-series in biomedical signals but are centered on finding the appropriate features for classification accuracy. While representing time-series of biomedical signals is important in developing an efficient and effective classification algorithm, such systems often have problems with the wide variety of variable signals necessary in a case-based reasoning system that wants to find similar but not exact matches. Work in this paper looks at identifying these similar signals in a large database efficiently.

2.2. Case-based reasoning

Many complexities in case-based diagnoses systems need to be addressed as datasets grow. Indeed, as patients across various locations submit data through systems, the potential for heterogeneous datasets becomes a problem. Work in [8] highlights an important missing data solution in order to present a comprehensive dataset for analysis. Such a system will present even larger data for use and searches of this size of data are those which this paper is concerned. Once datasets are completed, authors in [23] use medical signals for case-based reasoning, similar to the intended goal of this work. The system developed uses a k-NN clustering algorithm with Euclidean distance (R-NN based) for matching. Such systems, however, will result in a complex running time, will prune inefficient results, and/or will find too small a results set. Heuristics can be employed to improve the searches, such as by authors in [24] where a heuristic was developed to improve search for thyroid cancer treatment purposes. As will be shown in this work, the Monte Carlo method presented will result in greater information gain for such medical systems as in [23] using the idea of improving searches such as in [24], presenting improved results necessary for big data environments without increasing memory or computational complexities, shown both theoretically and empirically. This work will consider the exact R-NN matching scenario of previous works and show not only information gain improvement but also real-time performance gains.