Journal of Biomedical Informatics 42 (2009) 365–376

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Efficient layered density-based clustering of categorical data

Bill Andreopoulos [a,b,*], Aijun An [b], Xiaogang Wang [c], Dirk Labudde [a]

[a] *Biotechnological Centre, Technische Universität Dresden, 47-51 Tatzberg, 01307 Dresden Sachsen, Germany*
[b] *Dept. of Computer Science and Engineering, York University, Toronto, Canada*
[c] *Dept. of Mathematics and Statistics, York University, Toronto, Canada*

ARTICLE INFO

Article history:
Received 27 June 2008
Available online 10 December 2008

Keywords:
Clustering
Bioinformatics
Categorical
Network
Index
Scalable
Layered

ABSTRACT

A challenge involved in applying density-based clustering to categorical biomedical data is that the ”cube” of attribute values has no ordering defined, making the search for dense subspaces slow. We propose the HIERDENC algorithm for *hierarchical density-based clustering of categorical data*, and a complementary index for searching for dense subspaces efficiently. The HIERDENC index is updated when new objects are introduced, such that clustering does not need to be repeated on all objects. The updating and cluster retrieval are efficient. Comparisons with several other clustering algorithms showed that on large datasets HIERDENC achieved better runtime scalability on the number of objects, as well as cluster quality. By fast collapsing the bicliques in large networks we achieved an edge reduction of as much as 86.5%. HIERDENC is suitable for large and quickly growing datasets, since it is independent of object ordering, does not require re-clustering when new data emerges, and requires no user-specified input parameters.

1. Introduction

Categorical datasets are frequently clustered in biomedical informatics. Applications range from health information records to protein–protein interaction and sequence similarity networks. Layered categorical clustering, where a cluster consists of a center of similar objects and outer layers of less similar objects, has acquired prominence. Layered clusters are useful in bioinformatics for finding protein modules, complexes, and for visualization purposes [1–5]. However, often, the focus is on the quality of the clusters, with a secondary priority placed on the speed of the method, scalability to large datasets, and its usability.

A *categorical* dataset with m attributes is viewed as an m-dimensional ”cube”, offering a spatial density basis for clustering. A cell of the cube is mapped to the number of objects having values equal to its coordinates [6]. Clusters in such a cube are regarded as *subspaces* of high object density and are separated by subspaces of low object density [7]. *Density-based* clustering algorithms, such as DBSCAN [8] or OPTICS [9], search for dense subspaces. A dense subspace is defined by a *radius* of maximum distance from a central point, and it has to contain many objects according to a threshold criterion [10]. With the radius gradually increasing to allow more objects in clusters, *layered clusters* result. Our goal is to tackle some of the general challenges of existing clustering approaches:

(i) The density of a subspace is often defined relative to a user-specified *radius* [1]. However, different radii are preferable for different subspaces of the cube [9]. In dense subspaces where no information should be missed, the search is more accurately done ‘cell by cell’ with a low radius of 1. In sparse subspaces a higher radius may be preferable to aggregate information.
(ii) The time requirement is often a problem in density-based clustering, since it may be too slow to find the densest subspace in a high-dimensional dataset, and the dataset may change often. In particular, since there is no ordering of attribute values, the cube cells have no ordering either. The search for dense subspaces could have to consider several orderings of each dimension of the cube to identify the best clustering [11–13].
(iii) Other challenges include: re-clustering needed when new objects are introduced, difficulty finding clusters within clusters, sensitivity to order of object input, or user-specified input parameters required with wrong values affecting the end result [14–17].

We present the *HIERDENC* algorithm for “hierarchical density-based clustering of categorical data”, which addresses the above challenges. HIERDENC clusters the m-dimensional cube representing the spatial density of a set of objects with m categorical attributes. To find its dense subspaces, HIERDENC considers an

* Corresponding author. Address: Biotechnological Centre, Technische Universität Dresden, 47-51 Tatzberg, 01307 Dresden Sachsen, Germany.
E-mail address: williama@biotec.tu-dresden.de (B. Andreopoulos).

[1] Although the term ‘radius’ is borrowed from geometrical analogies that assume circular constructs, we use the term in a looser way and it is not a Euclidean distance.

1532-0464/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2008.11.004

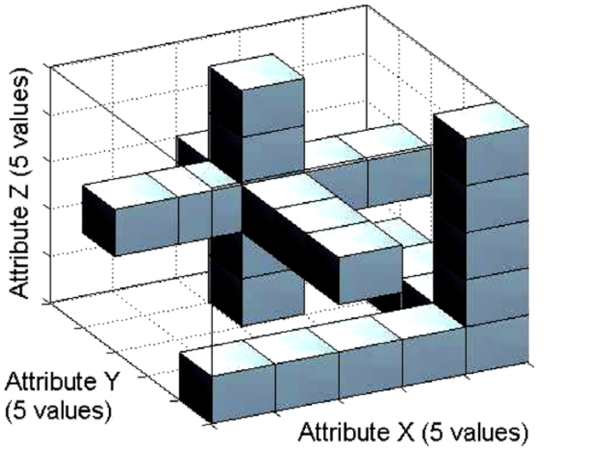

Fig. 1. Two HIERDENC 'hyper-cubes' for radius r=1, in a 3D cube. All neighbors of the central object for each hyper-cube differ from it in one dimension.

object's neighbors to be all objects that are within a radius of maximum dissimilarity. Fig. 1 shows that the radius is the maximum number of dimensions by which neighbors can differ. The cube search starts from a low radius and gradually moves to higher radii. With the clustering radius gradually increasing, *layered clusters* result, as Fig. 2 shows. Fig. 3 shows examples of creating and expanding clusters in a 3-dimensional dataset.

For scalability to large categorical datasets, we propose the *HIERDENC index*, which supports efficient retrieval of dense subspaces relative to a radius. When new objects are introduced, HIERDENC is updated efficiently. The neighborhood of an object is insensitive to attribute or value ordering. A user can study the layered cluster structure at different levels of granularity, detect subclusters within clusters, and know the central densest area of each cluster.

Applications of HIERDENC to biomedical informatics abound. One application is clustering networks to find bicliques. A network's adjacency matrix is a boolean-valued categorical dataset, where rows and columns represent objects and '1' is a connection; a biclique is a network whose objects can be divided into two disjoint sets, such that every object of the first set is connected to every object of the second set. Finding bicliques in a network has several applications to biological problems; in protein–protein interaction networks the bicliques can be visualized, or correlated with structural knowledge to find the structures that induce observed interactions[1,5]. Furthermore, HIERDENC is applicable to biomedical images and literature; we demonstrate a fast image retrieval system and a PubMed document clustering that we built. We also applied HIERDENC to clustering of Force–Distance curves from high-throughput proteomic studies, by aligning curves on the basis of their detected peaks to one another.

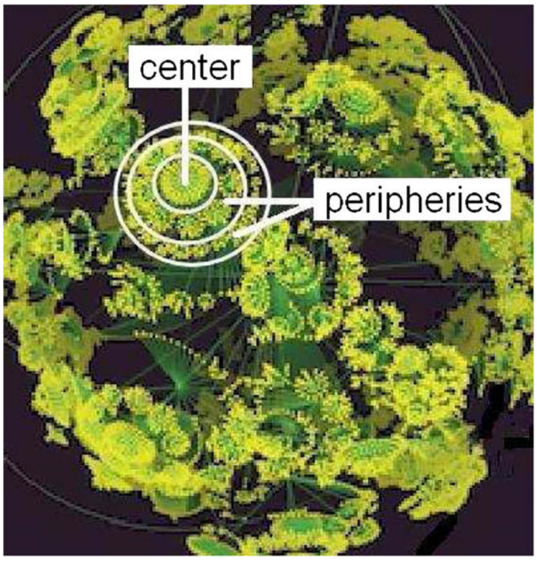

Fig. 2. A layered network cluster has a center surrounded by outer layers.

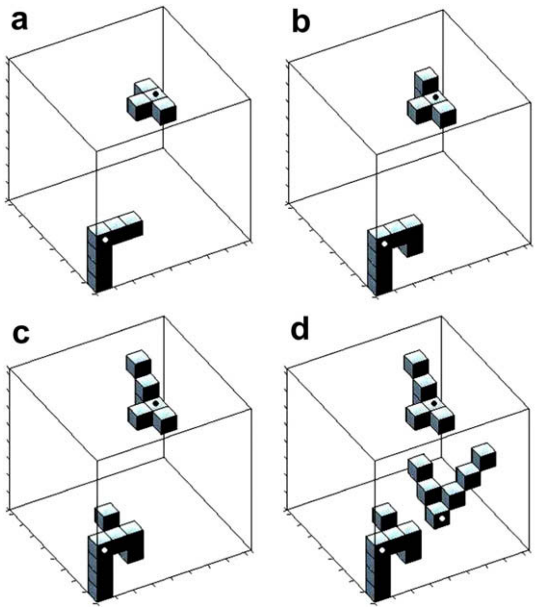

Fig. 3. A cluster is a dense subspace with a 'central' cell marked with a dot. The radius starts from 1 and changes when neither a cluster can be expanded, nor a new cluster can be formed. (a) r=1, two new clusters. (b) r=1, clusters expand. The radius did not change since *a*, but fewer objects are found within a radius of 1 than in *a*, implying a less dense subspace. (c) r=2, clusters expand. (d) r=2, one new cluster.

This paper is organized as follows. Section 2 gives an overview of related work on dissimilarity metrics and density-based clustering for categorical data. Sections 3 and 4 present the HIERDENC clustering algorithm and index. Section 5 discusses our performance evaluations. Applications to categorical datasets show runtime scalability and clustering quality. In Section 6, we discuss biomedical applications. We apply HIERDENC to large networks, such that bicliques are collapsed, comfirming the runtime scalability. Section 7 concludes that our method amends some of the weaknesses of previous categorical density-based clustering approaches, and has promising utilities in biomedical informatics.

2. Background and related work

Section 2.1 describes the Hamming distance, and Section 2.2 provides an overview of density-based clustering algorithms for categorical data.

2.1. The Hamming distance in categorical and binary data

For a fixed length m, the Hamming distance is a metric on the vector space of the words of that length. Fig. 4 shows an example of HDs in the *zoo* dataset [18]. The serpent *tuatara* is within a relatively small HD from the other serpents; the maximum distance is $HD(tuatara \leftrightarrow seasnake) = 5$. On the other hand, $HD(tuatara \leftrightarrow gorilla) = 8$, and gorilla is unlikely to belong to the class of serpents. For binary strings a and b the HD is equivalent to the number of ones in a *xor* b. The metric space of binary strings of length m, together with the HD metric, is known as the Hamming cube.