Information Sciences 370–371 (2016) 343–354

Contents lists available at ScienceDirect

Information Sciences

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

A-Ward$_{p\beta}$: Effective hierarchical clustering using the Minkowski metric and a fast k-means initialisation

Renato Cordeiro de Amorim[a,*], Vladimir Makarenkov[b], Boris Mirkin[c,d]

[a] School of Computer Science, University of Hertfordshire, College Lane Campus, Hatfield AL10 9AB, UK
[b] Département d'Informatique, Université du Québec à Montréal, C.P. 8888 succ. Centre-Ville, Montreal (QC) H3C 3P8, Canada
[c] Department of Data Analysis and Machine Intelligence, National Research University Higher School of Economics, Moscow, Russian Federation
[d] Department of Computer Science and Information Systems, Birkbeck University of London, Malet Street, London WC1E 7HX, UK

ARTICLE INFO

Article history:
Received 28 March 2016
Revised 19 July 2016
Accepted 31 July 2016
Available online 1 August 2016

Keywords:
Initialisation algorithm
Minkowski metric
Hierarchical clustering
Feature weighting

ABSTRACT

In this paper we make two novel contributions to hierarchical clustering. First, we introduce an anomalous pattern initialisation method for hierarchical clustering algorithms, called A-Ward, capable of substantially reducing the time they take to converge. This method generates an initial partition with a sufficiently large number of clusters. This allows the cluster merging process to start from this partition rather than from a trivial partition composed solely of singletons.

Our second contribution is an extension of the Ward and Ward$_p$ algorithms to the situation where the feature weight exponent can differ from the exponent of the Minkowski distance. This new method, called A-Ward$_{p\beta}$, is able to generate a much wider variety of clustering solutions. We also demonstrate that its parameters can be estimated reasonably well by using a cluster validity index.

We perform numerous experiments using data sets with two types of noise, insertion of noise features and blurring within-cluster values of some features. These experiments allow us to conclude: (i) our anomalous pattern initialisation method does indeed reduce the time a hierarchical clustering algorithm takes to complete, without negatively impacting its cluster recovery ability; (ii) A-Ward$_{p\beta}$ provides better cluster recovery than both Ward and Ward$_p$.

1. Introduction

Clustering algorithms are a popular choice when tackling problems requiring exploratory data analysis. In this scenario, analysts can draw conclusions about data at hand without having information regarding the class membership of the given entities. Clustering algorithms aim at partitioning a given data set Y into K homogeneous clusters $S = \{S_1, S_2, \ldots, S_K\}$ without requiring any label learning process. These algorithms summarise information about each cluster by producing K centroids, often called prototypes, $C = \{c_1, c_2, \ldots, c_K\}$. The ability to partition data and to provide information about each part has made the application of clustering popular in many fields, including: data mining, computer vision, security, and bioinformatics [17,20,23,24,26,35].

[*] Corresponding author.Fax: +44 01707 284115.
E-mail addresses: r.amorim@herts.ac.uk (R. Cordeiro de Amorim), makarenkov.vladimir@uqam.ca (V. Makarenkov), bmirkin@hse.ru (B. Mirkin).

http://dx.doi.org/10.1016/j.ins.2016.07.076
0020-0255/© 2016 Elsevier Inc. All rights reserved.

There are various approaches to data clustering, with algorithms often divided into partitional and hierarchical. Originally, partitional algorithms produced only disjoint clusters so that each entity $y_i \in Y$ was assigned to a single cluster S_k. This hard clustering approach has been variously extended to fuzzy sets [42]. Fuzzy clustering allows a given entity $y_i \in Y$ to belong to each cluster $S_k \in S$ with different degrees of membership. There are indeed a number of partitional algorithms, with k-means [2,22] and fuzzy c-means [3] being arguably the most popular under the hard and fuzzy approach, respectively.

Hierarchical algorithms provide additional information about data. They generate a clustering S and related set of centroids C, very much like partitional algorithms, but they also give information regarding the relationships among clusters. This information comes as a nested sequence of partitions. This tree-like relationship can be visualized with a dendrogram (i.e., an ultrametric tree). In this type of clustering, an entity $y_i \in Y$ may be assigned to more than one cluster as long as the clusters are related and the assignment occurs at different levels of the hierarchy.

Hierarchical algorithms can be divided into agglomerative and divisive [26]. Agglomerative algorithms follow a bottom-up approach. They start by setting each entity $y_i \in Y$ as the centroid of its own cluster (singleton). Pairs of clusters are then merged stepwise until all the entities have been collected in the same cluster, or until a pre-specified number of clusters is found. Divisive algorithms do the opposite by following a top-down approach.

There is indeed a wide variety of algorithms to apply when using hierarchical clustering. The Ward method [39] is one of the most popular hierarchical algorithms. It follows the agglomerative approach, merging at each iteration the two clusters that minimise the within-cluster variance. This variance is measured as a weighted sum of squares, taking into account the cardinality of each cluster, and leading to the cost function as follows:

$$Ward(S_a, S_b) = \frac{N_a N_b}{N_a + N_b} \sum_{v=1}^{V} (c_{av} - c_{bv})^2, \tag{1}$$

where V is the number of features used to describe each entity $y_i \in Y$. N_a and c_a represent the cardinality and centroid of cluster $S_a \in S$, respectively. Similarly, we have N_b and c_b for cluster $S_b \in S$. The fraction in (1) ensures that if two pairs of clusters are equally apart, those of lower cardinalities are merged first.

Previously, we extended the traditional Ward algorithm by introducing Ward$_p$ [8]. Our algorithm applies cluster dependent feature weights and extends the squared Euclidean distance in (1) to the p-th power of the weighted Minkowski distance. With these we: (i) ensure that relevant features have a higher impact in the clustering than those that are less relevant; (ii) can set the distance bias to other shapes than that of a spherical cluster, a problem traditionally addressed by methods following model-based clustering [14].

The contribution of this paper is two-fold. First, we introduce what we believe to be the first non-trivial initialisation method for a hierarchical clustering algorithm. Our method generates an initial partition with a sufficiently large number of clusters. Then, the merging process applies starting from this partition rather than from the singletons. In this way, the running time of a given hierarchical clustering algorithm is substantially reduced. Second, we advance hierarchical clustering by introducing A-Ward$_{p\beta}$, an extension of Ward$_p$ to the situation in which our initialisation method applies and the feature weight exponent can differ from the exponent of the Minkowski distance. We give a rule for choosing these two exponents for any given data set. We run numerous computational experiments, with and without noise in data sets.

It is worth noting that the “noise” in this paper has nothing to do with the conventional meaning of measurement errors, which are usually modelled by an additive or multiplicative Gaussian distribution affecting every data entry. Here, the noise is modelled by either of two ways: (1) inserting additional random noise features, and (2) blurring some features within some clusters. We establish that: (i) the initial clustering generated by our method does decrease the time a hierarchical clustering algorithm takes to complete; (ii) A-Ward$_{p\beta}$ provides a better cluster recovery under different noise models, than either the Ward or the Ward$_p$ algorithms, especially for noisy data.

We direct readers interested to know more of feature weighting in the square-error clustering to reviews such as [19], and references within.

2. Ward clustering using anomalous patterns

2.1. Ward and anomalous pattern Ward

K-means is arguably the most popular partitional clustering algorithm [17,35]. It can be considered an analogue to the general expectation-maximisation algorithm (EM) [12]. Note, however, that EM recovers a mixed distribution density function, whereas k-means just finds a set of non-overlapping clusters and their centres. K-means alternatingly minimises the within cluster sum of squares:

$$W(S, C) = \sum_{k=1}^{K} \sum_{y_i \in S_k} \sum_{v=1}^{V} (y_{iv} - c_{kv})^2 \tag{2}$$

to obtain a partition of the given set of N entities in a set of non-overlapping clusters $S_k \in S$, each represented by its centroid c_k, $k = 1, 2, \ldots, K$. This minimisation is usually done by following the three straightforward steps: (i) set the coordinates of each centroid $c_k \in C$ to a randomly chosen entity $y_i \in Y$; (ii) assign each entity $y_i \in Y$ to the cluster S_k whose centroid c_k is