Clustering-Based Data Cleaning Postprint for Solar Photosphere Bright Points

Authors: Zhang Aili, Xiong Jianping, Yang Yunfei, Feng Song, Deng Hui, Ji Kaifan

Date: 2017-10-20T00:00:00+00:00

Abstract

Due to the small scale of photospheric bright points and their indistinct edge structures, some bright granules are inevitably misidentified as bright points during recognition. Partition-based K-means algorithm and density-based DBSCAN algorithm are employed to clean the feature data of all bright structures, aiming to eliminate non-bright-point structures from bright-point structures. First, the LMD algorithm and the concept of 3D connectivity are used to identify and track bright points, and then seven relatively uncorrelated feature values of the bright points are extracted, including equivalent diameter, intensity, eccentricity, the proportion of bright point edges located in intergranular dark lanes, velocity, motion pattern, and diffusion coefficient. After data standardization, Principal Component Analysis (PCA) is applied to reduce the dimensionality to three dimensions based on a 90% contribution rate. Finally, K-means algorithm and DBSCAN algorithm are used to clean the bright point data. Experimental results show that both algorithms can clean non-bright-point structures, with the accuracy of the K-means algorithm being 80% and that of the DBSCAN algorithm being 53%. Therefore, the K-means algorithm can more effectively distinguish between bright-point and non-bright-point structures.

Full Text

Data Cleaning of Solar Photospheric Bright Points Based on Clustering

Zhang Aili, Xiong Jianping, Yang Yunfei, Feng Hui, Ji Kaifan
(Yunnan Key Laboratory of Computer Technology Application, Kunming University of Science and Technology, Kunming, Yunnan 650500, China, Email: jikaifan@cnlab.net)

Abstract

Due to their small scale and indistinct boundary structures, a portion of bright granular fragments are inevitably misidentified as photospheric bright points (PBPs) during detection. This study employs partition-based clustering algorithms to clean the feature data of all bright structures, aiming to separate non-PBP structures from true PBPs. First, a three-dimensional connectivity approach is used to identify and track bright points, extracting seven relatively independent feature values including equivalent diameter, intensity, eccentricity, proportion of bright point boundaries located in intergranular dark lanes, velocity, motion type, and diffusion coefficient. After data standardization, principal component analysis (PCA) is applied to reduce the dimensionality to three dimensions based on contribution rates. Finally, both K-means and density-based DBSCAN algorithms are used to clean the bright point data. Experimental results demonstrate that both algorithms can effectively remove non-PBP structures, with the K-means algorithm achieving a higher accuracy rate than DBSCAN. The K-means algorithm thus proves more effective at distinguishing between PBPs and non-PBP structures.

Keywords: Photospheric bright points; Non-bright point structures; Clustering algorithm; K-means algorithm; DBSCAN algorithm

1. Introduction

The solar photosphere is covered with granular structures. Some bright structures appear in the dark lanes between granules, known as photospheric bright points (PBPs). PBPs are closely related to magnetic fields, and studying them can advance our understanding of solar magnetism, deeper and hotter plasma processes, and coronal heating phenomena. However, PBPs are easily confused with bright granular fragments and other small-scale features with locally high intensity on the solar surface.

Current two-dimensional identification methods primarily employ thresholding, region growing, and morphological techniques. Thresholding divides image grayscale levels by setting one or more thresholds, treating pixels within the same range as belonging to the same object. Region growing starts from an initial region and gradually merges adjacent pixels or regions with similar properties until no more merges are possible. Morphology uses structural elements of specific shapes to measure and extract corresponding shapes from images for analysis and recognition. However, these methods inevitably misidentify some bright granular fragments as PBPs.

Data cleaning is an emerging technology that has developed alongside data mining, involving the detection and correction of erroneous and inconsistent data from datasets to improve data quality. In recent years, scholars have proposed using clustering methods for data cleaning. Clustering analysis is a statisti-

cal technique that groups research objects into relatively homogeneous clusters, placing similar objects in the same category and dissimilar ones in different categories. Clustering algorithms can be categorized into partitioning methods, hierarchical methods, density-based methods, grid-based methods, and model-based methods. Partitioning and density-based methods are particularly efficient for large datasets.

This paper proposes using clustering analysis to clean PBP data, achieving the goal of separating non-PBP structures from PBP candidates. Section 1 describes the data source and extraction methods. Section 2 introduces the data preprocessing and clustering approaches. Section 3 presents the cleaning results and analysis. Section 4 provides a summary and conclusions.

1.1 Data Source and Extraction

The experimental data were obtained from the Hinode/Solar Optical Telescope (SOT; Ichimoto et al. 2004; Suematsu et al. 2008) G-band observations of a quiet region near disk center. The pixel resolution is 0.054 arcsec/pixel, the field of view is 20 arcsec $\times$ 20 arcsec, and the temporal resolution is 11 seconds. First, a local correlation-based sub-pixel alignment algorithm was used to align the image sequence. The Laplacian and Morphological Dilation (LMD) algorithm was then employed to identify photospheric bright points.

[Figure 1: see original paper] shows the G-band image observed by Hinode/SOT at 18:19 UT on February 19, 2007, and the PBPs detected by the LMD algorithm. After identifying bright points in each frame of the sequence, a three-dimensional spatiotemporal cube approach was used to track PBPs. A bright point that does not undergo merging or splitting during its lifetime is called an isolated point; otherwise, it is non-isolated. The evolution of an isolated point appears as a cylindrical structure, with its horizontal velocity displayed as distortion of this cylinder along the time axis, and its lifetime representing the temporal extent of the cylinder.

1.2 Feature Extraction for PBP Data

Seven relatively independent features were selected to characterize PBPs: equivalent diameter, intensity, eccentricity, proportion of boundary in dark lanes, velocity, motion type, and diffusion coefficient. These features have low correlation and can represent the optical, shape, and motion characteristics of PBPs.

The definitions are as follows: - **Equivalent diameter**: The area of all pixels belonging to a bright point is calculated and converted to an equivalent circle diameter. - **Intensity**: The maximum intensity of a bright point divided by the average intensity of the entire image. - **Eccentricity**: The distance between the two foci of an ellipse divided by its major axis length. Higher eccentricity indicates a more elongated shape, while lower values indicate a more circular shape. - **Proportion of boundary in dark lanes**: Since PBPs are typically located in intergranular dark lanes, this feature extracts the proportion of each

bright point' s boundary that lies within dark lanes. - **Velocity**: Calculated from the displacement of a bright point' s centroid between consecutive frames. - **Motion type**: Defined by parameter mt, which quantitatively describes the trajectory. mt is calculated as the ratio of total displacement to trajectory length. Values range from 0 to 1, where mt approaching 1 indicates linear motion, and values near 0 indicate circular trajectories that return to the origin. - **Diffusion coefficient**: Describes the relationship between diffusion area and time. The mean square displacement between a point' s position at any time and its initial position is given by $\langle(\Delta r)^2\rangle = 4DT^\gamma$, where D is the diffusion coefficient and T is the lifetime. Larger diffusion coefficients indicate greater area diffused per unit time.

Since each bright point has multiple attribute values during its lifetime, we calculate the average values for each attribute (e.g., average diameter, average eccentricity, average boundary proportion, average velocity) to represent its entire evolutionary characteristics.

2. Data Preprocessing

2.1 Data Standardization Because the seven attributes have different units and scales, standardization is required to convert them into dimensionless values. Standardization removes data units, enabling comparison and weighting of indicators with different units or magnitudes. The z-score method is used, where μ is the mean of all sample data and σ is the standard deviation. After standardization, the data follow a standard normal distribution.

2.2 Principal Component Analysis High-dimensional data contains substantial redundant information, so dimensionality reduction is necessary. PCA is an unsupervised feature extraction method that uses linear transformation to convert multi-dimensional features into fewer dimensions while preserving most of the original information. The process involves: (1) constructing a feature matrix from sample data, (2) calculating the covariance matrix to reveal relationships between dimensions, (3) computing eigenvectors and eigenvalues, (4) sorting eigenvalues in descending order, and (5) selecting the top k components based on contribution rates.

The contribution rate indicates the proportion of variance explained by each principal component. When the cumulative contribution rate of the first k components reaches a satisfactory level, they can reliably replace the original variables. For this dataset, the cumulative contribution rate reaches over 80% by the third principal component, indicating that three dimensions can adequately represent the original seven-dimensional data.

[Figure 2: see original paper] shows the relationship between contribution rate and principal components.

2.2 Clustering Algorithms

2.2.1 K-means Algorithm K-means is a representative partitioning clustering algorithm that groups n objects into k clusters based on nearest distance principles. The algorithm randomly selects k initial centroids, then iteratively assigns each object to the nearest centroid and recalculates cluster means until convergence. The objective function minimizes within-cluster variance:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \|x_i^{(j)} - c_j\|^2$$

where $x_i^{(j)}$ represents data objects and c_j represents cluster centroids. The goal is to produce compact, well-separated clusters.

2.2.2 DBSCAN Algorithm DBSCAN is a density-based clustering algorithm that discovers clusters of arbitrary shape by identifying high-density regions. Key definitions include: - **-neighborhood**: The region within radius of a given object. - **Core object**: An object whose -neighborhood contains at least MinPts objects. - **Directly density-reachable**: Object p is directly density-reachable from q if p is within q' s -neighborhood and q is a core object. - **Density-reachable**: Object p is density-reachable from q if there exists a chain of objects where each is directly density-reachable from the previous one. - **Density-connected**: Objects p and q are density-connected if there exists an object O from which both are density-reachable.

The algorithm processes as follows: for each point, if its -neighborhood contains more than MinPts points, a new cluster is created with that point as a core object. The cluster is then expanded by adding all density-reachable objects. The process continues until no new points can be added to any cluster.

3. Clustering Results

3.1 K-means Algorithm Results The K-means algorithm was applied with the target cluster number set to 2 (representing PBPs and noise points). However, visual inspection of the original images revealed too many noise points being identified. [Figure 3: see original paper] shows the cleaning results displayed in a 2D image, where blue represents PBPs and red represents noise points.

[Figure 4: see original paper] displays the three-dimensional spatiotemporal evolution structures, with blue indicating PBPs and red indicating noise points. The noise structures show varied lengths and trajectories. Further analysis in the time series [Figure 5: see original paper] illustrates the evolution of both PBPs and noise points, with three different scenarios marked. Position (a) shows a noise point that remains on granules throughout its lifetime, position (b) shows a PBP that remains in intergranular dark lanes, and position (c)

reveals a case where the algorithm misclassified a structure that appears on granules at 19:02:50 UT, indicating cleaning errors.

Through comprehensive analysis of all evolutionary tracks, 80% of structures identified as non-PBPs by K-means were confirmed to be true non-PBPs.

3.2 DBSCAN Algorithm Results The DBSCAN algorithm results are shown in [Figure 6: see original paper] (2D display) and [Figure 7: see original paper] (3D spatiotemporal cube). Similar to the K-means analysis, [Figure 8: see original paper] shows the time series evolution for DBSCAN-cleaned data. The analysis revealed that 53% of structures identified as non-PBPs by DBSCAN were confirmed to be true non-PBPs, with the remaining 47% representing classification errors.

4. Summary and Outlook

This paper employs clustering methods to clean PBP data, successfully separating non-PBP structures from PBP candidates. The process involves LMD algorithm identification, three-dimensional spatiotemporal tracking, extraction of seven representative features, z-score standardization, PCA dimensionality reduction to three dimensions (with cumulative contribution rate >80%), and finally clustering using K-means and DBSCAN algorithms.

Both algorithms can clean non-PBP structures, with K-means achieving approximately 80% accuracy compared to DBSCAN' s 53%. This indicates K-means is more suitable for cleaning non-PBP structures. This method provides an effective approach for removing inevitable noise in PBP identification, yielding more accurate data for small-scale magnetic field studies and coronal heating research.

Future improvements should address the current limitations: reducing classification errors, decreasing parameter dependency, and considering physical parameters such as spatial resolution, which may affect features like eccentricity. Different observational resolutions may require adjusted parameters and weights to achieve more accurate cleaning results.

References

[1] Jess D B, et al. Magnetic bright points in the quiet Sun. *The Astrophysical Journal Letters*, 1852-1861.

[2] Wang Yongmei, Chen Jiaqi, Geng Yuliang. An interactive data cleaning system. *Computer Engineering and Design*, 955-957.

[3] Liu Yanxiao, Yang Yunfei, Lin Jun. A region-growing algorithm to recognize magnetic bright spots in the solar photosphere. *Astronomical Research & Technology—Publications of National Astronomical Observatories of China*, 145-150.

[4] Almeida J S, Bonet J A, et al. Magnetic bright points in the quiet Sun. *The Astrophysical Journal Letters*, L26-L29.

[5] Bovelet B, Wiehr E. Multiple-scale pattern recognition applied to faint intergranular G-band structures. *Solar Physics*, 121-129.

[6] Crockett P J, Mathioudakis M, et al. Automated detection and tracking of solar active region and quiet Sun. *Monthly Notices of the Royal Astronomical Society*, 201-206.

[7] Chen Jiaqi, Geng Yuliang. Interactive data cleaning system. *Computer Engineering and Design*, 955-957.

[8] Guo Zhimao, Zhou Aoying. Research on data quality and data cleaning: A survey. *Journal of Software*, 2076-2082.

[9] Sun Jigui, Liu Jie, Zhao Lianyu. Study on clustering algorithms. *Journal of Software*, 48-61.

[10] Wunsch D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 645-678.

[11] Chen Jie, Feng Song, Deng Hui, et al. Comparison of correlation-based techniques for correcting and stacking solar magnetic-field images. *Astronomical Research & Technology*, 17-24.

[12] Qu Huixue, Ji Kaifan, et al. Characterizing motion types of G-band bright points in the quiet Sun. *Research in Astronomy & Astrophysics*, 569-582.

[13] Tranchevent L C, Moor B D, et al. Optimized data fusion for kernel k-means clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 209-217.

[14] Yang Yunfei, et al. Statistical study of photospheric bright points in the quiet Sun. *Astrophysics and Space Science*, 1852-1861.

[15] Ghosh S, Dubey S K. Comparative analysis of k-means and fuzzy c-means algorithms. *International Journal of Advanced Computer Science and Applications*, 35-39.

[16] Patel B C. Adaptive k-means clustering algorithm for breast image segmentation. *International Journal of Computer Applications*, 35-38.

[17] Sinha D G R. An adaptive k-means clustering algorithm for breast image segmentation. *IJACSA*.

[18] Zhou Aoying, Zhou Shuiheng, Cao Jing, et al. Approaches for scaling DBSCAN algorithm to large spatial databases. *Journal of Computer Science and Technology*, 509-526.

Note: Figure translations are in progress. See original paper for figures.