International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Survey on Hubness - Based Clustering Algorithms

Nikita Dhama[1], Antara Bhatttacharya[2]

[1, 2]Department of CSE, G H Raisoni Institute of Engineering and Technology for Women RTMNU, Nagpur, MH, India

Abstract: ***Clustering of high dimensionality data which can be seen in almost all fields these days is becoming very tedious process. The key disadvantage of high dimensional data which we can pen down is curse of dimensionality. As the magnitude of datasets grows the data points become sparse and density of area becomes less making it difficult to cluster that data which further reduces the performance of traditional algorithms used for clustering. To rout these toils hubness based algorithms were introduced as a variation to these algorithms which influences the distribution of the data points among the k-nearest neighbor. The hubness is an unguided method which finds out which points appear more frequently in the k-nearest neighbor than other points in the dataset. This paper discusses the ways of clustering algorithms using hubness phenomenon. One of the methods is based on condensed nearest neighbor who is performed iteratively on the order independent data. The next algorithm is hinged for fuzzy a based approach which performs better on uncertain data i.e. partially exposed or incomplete data. The proposed algorithms are basically used for increasing the efficiency and increasing predicting accuracy of the system.***

Keywords: clustering, high dimensional data, hubness, nearest neighbour

1. Introduction

Clustering of data provides us with a way to group elements together such that elements of same group are of similar attributes or features. Based on the enactment of clusters the criteria for clustering changes. Clustering is often muddled with classification, but classification differs with clustering in a way that in clustering both classes and the objects included in clustering are already defined i.e. predefined. With the help of clustering techniques objects which are logically similar to each other are physically kept near to each other. According to [1] the various clustering algorithms are randomly sketched into 4 types namely: partitional algorithms, hierarchical algorithms, density based algorithms, and subspace algorithms. Out of these the subspace algorithms are basically used to cluster high dimensionality data. High dimensionality usually refers to a large number of attributes of the specified objects. When the dimension of the data increases it leads to the curse of dimensionality which reduces the performance of the clustering algorithms. The curse of dimensionality refers to the problem of handling the data when the number of dimensions increases.

The concept of hubness is used to handle datasets containing high dimensional data points. Due to the increasing dimensions of a data set sharing of the number of times a data point appears among the *k* nearest neighbors of other data points in the datasets becomes increasingly bevelled. As the dimensions of the data points increases, the time needed for the execution increases and efficiency required goes on decreasing. The traditional machine learning algorithms and methods can be further modified to increase the accuracy and the efficiency of algorithms. The algorithms to be used are based on the k nearest neighbor technique of clustering. The shared neighbor algorithm will support to relate data of different clusters which in turn will provide better clustering. Density based approaches can also be added with the shared neighbor algorithm to provide more efficiency.

The fast condensed nearest neighbour algorithm is an advancement to condense nearest neighbour algorithm which is used to reduce the dataset for classification based on some prototypes. Here the learning speed of the algorithm is also considered which tells us that the algorithm should give good behaviour under all conditions. The Fuzzy Rough Nearest Neighbor algorithm is based on the classification that the the description of vector space is not proper. Due to this we get imperfect classification space resulting in the uncertainty of result. The FRNN algorithm can provide us with more accurate prediction of clustering result. The k-NN algorithm gives results which are satisfactory for high dimensional data. The applications of these clustering algorithms can be seen in various fields like text mining, text retrieval, classification image feature and many more.

The remainder of this paper is organized as follows: Section 2 surveys the different clustering algorithm and its types. Section Section 3 presents a brief review of proposed methods Section 4 examines the conclusion of the proposed propagation model.

2. Related Work

In what follows, the related work on mobile virus and their propagation models is reviewed first. Next, some virus defense methods that contain abnormal detection technologies for restraining virus propagation in mobile networks are introduced here.

A. Clustering algorithms

Clustering is the process of grouping data elements based on some specific properties. Different algorithm considers different algorithms while clustering the datasets. The choice of algorithm for clustering depends upon the datasets on which we want to apply the algorithm. From [1] four rough categories of clustering are defined. Hierarchical clustering algorithms occur in iteration by either merging smaller clusters in to larger ones or dividing larger clusters to smaller ones. Whichever way, it forms a hierarchy of clusters usually known as a dendogram. Partitional clustering algorithms form various partitions and then calculate them by some conditions. They are sometimes called as nonhierarchical as each occurrence is placed in exactly one of k mutually exclusive clusters. Density based algorithms for clustering take clusters from dense region of datasets rather than the border regions where lot of noise is found. But when the

Licensed Under Creative Commons Attribution CC BY

dimensionality of data sets increases then the sparsity of data elements increases. So for these types of data sets we can use the fourth type of clustering algorithm i.e. subspace clustering algorithms. In these subsets of the data sets are formed and then clustering is applied to it. This helps to increase efficiency of clustering high dimensional data.

Table 1: Types of Clustering Algorithms

Clustering algorithms	*Examples*
Partitional	K-means, k-medoids.
Heirarchical	Agglomerative,divisive.
Density based	DBScan.
Subspace	k-nn,Scaf

B. Concept of hubness

Subspace algorithms can be implemented using the hubness concept. Hubness is the tendency of some data points in high-dimensional data sets to occur much more frequently in k-nearest neighbor lists of other points than the rest of the points from the set, can in fact be used for clustering. Hubs can be viewed as referent to outliers. The outliers have high inter- and intracluster distance, telling that hubs must also get separate attention. The various Consequences and applications of hubness have been more properly studied in other related fields like the classification image feature, representation, data reduction, collaborative filtering, text retrieval, and music retrieval. Hubs become a center of attraction within the data and a small number of points that determines most characteristics of the datasets. Less number of points can play an useful role while doing classification. These points may cause many label to mismatch in the k-nearest neighbor sets, so are also called 'bad hubs', while the remaining are called as 'good hubs'. The good hubs help to increase the quality of clustering while the bad hubs reduce the clustering quality. So algorithms try to reduce the number of bad hubs in the datasets to get better results.

C. Nearest Neighbour Algorithms

The nearest neighbor (NN) rule is a nonparametric method for pattern classification based on instances. Introduced by Fix and Hodges in 1951, the NN rule gained considerable popularity after 1967, when some of its formal properties were described by Cover and Hart. Cover's work was a milestone in a subject which has since become a lively research field for many researchers in pattern recognition and machine learning and the study and development of one of the top ten algorithms in data mining. The measurement of distances is another facet of curse of dimensionality, which is a common term for problems of clustering of datasets of high dimensional spaces. It is the fortuitous nature of all elements in a high dimensional space to be at the equal distance from remaining points in that space. It is usually calculated as a ratio between spread and magnitude i.e. the ratio between the ideal deviation of all distances to an random reference point and the mean of these distances. If the ideal deviation remains more or less constant with growing dimensionality while the mean keeps increasing, the ratio coincide to zero with dimensionality going to infinity.

In this work, we incorporate research on subspace clustering algorithms which provide more clustering efficiency. The main advantage of the proposed algorithms is their cluster efficiency on high dimensional data. The K-nearest neighbour algorithms are based on ways which do not use parameterised families of the probability used for classification or regression. They do not make assumption of distribution of elements. The fuzzy based approach is basically used when there is imbalance in domain or when the data is partially exposed. The condensed nearest neighbour algorithm is a a novel order independent algorithm for finding a training set consistent subset for the NN rule, called FCNN rule which is an hybrid approach. Hybrid methods search for a small subset of the training set that, at the same time it gains both noisy and redundant instances obliteration. Competence enhancement method and preservation methods are joined to gain the same aim of hybrid methods.

3. Methodology

Although various methods for clustering high dimensional data are available we still prefer hubness based algorithms when clustering high dimensional data sets. This is because of the fact that the hubness based algorithms are more efficient in clustering these data elements as compared to other cluttering algorithms. The proposed algorithms are based on the nearest neighbour algorithm methodology. Various types of algorithms are available which helps in clustering. Two such methods are defined below namely Fuzzy Rough NN, and Fast Condensed NN.

Table 2: Various algorithm comparison

Sr. No.	*Technique*	*Idea*	*Advantages*	*Target*
1.	K Nearest Neighbor	Uses nearest neighbor rule	1. training is very fast 2. Simple and easy to learn 3. Robust to noisy training data 4.Effective if training data is large	Large data samples
2.	Reduced Nearest Neigh (RNN)	Remove patterns which do not affect the training data set results	1. Reduce size of training data and eliminate templates 2. Improve query time and memory requirements 3.Reduce the recognition rate	Large data set
3.	Rank nearest neighbour	Assign ranks to training data for each category	1.Performs better when there are too much variations between features 2.Robust as based on rank	Class distribution of Gaussian nature
4.	Condensed nearest Neighbour	Eliminate data sets which show similarity and do not add extra information	1. Reduce size of training data 2. Improve query time and memory requirements 3.Reduce the recognition rate	Data set where memory requirement is main concern
5.	Fast condensed	Select points very close to the	1. Has a smaller complexity than CNN	

Licensed Under Creative Commons Attribution CC BY

	nearest neighbor	decision boundary	2. Good rate of condensation 3. Independent of the order	
6.	Fuzzy nearest neighbour	interpreting each entry in the database as a point in space	1. Reduce size of training data 2. Similarity measure: 3. Decision rule by voting means	Large datasets with feature detection

Recently, different ways to nearest neighbor classification based on fuzzy rough sets have been given. Most of them aim to improve the quality of the classification performed with the combined support of the rough sets and fuzzy sets theories. Proposal of FRNN, for the first time was presented in[11]. This classifier incorporates the lower and upper approximations of the memberships to the decision rule, in an attempt to deal with fuzzy vagueness and rough uncertainties. A second proposal of FRNN develops this facet further, associating fuzzy vagueness with the existing overlapping between classes and rough uncertainties with the lack of a proper number of features to describe the data. Another main feature of this method is that it does not require a fixed k value for the classification rule.

Fuzzy-rough nearest neighbor classification is developed in various research works where the characteristics of these are given. The FRNN works in two steps. In the first step for a test pattern it tries to find the k-NN and then use this k-NN to approximate the test pattern. They make use of fuzzy rough datasets and vaguely quantified rough sets. Then the first classifier can be represented as an advancement of FRNN, whereas in the next one we can then vaguely quantified rough sets that are introduced to reduce the responsiveness of the classifier to noise Qu et al. presents an approach to hybridizing kernel-based classification which uses vornoi diagram, with the fuzzy rough sets.

Data points in the training datasets are divided in:

1) **Outliers**: points that are not recognised as the correct type if given to the database afterwards.
2) **Prototypes**: based on which clustering is done and are needed in training dataset for other non-outlier elements to be accurately recognised.
3) **Absorbed points**: points which are not outliers, and would be correctly recognised based just on the set of prototype points New elements are only needed to be compared with the prototype points, instead of the complete database.

After applying data reduction, we can classify new samples by using the k-NN algorithm against the set of prototypes. Classifying a new sample data points is now rapid, since we don't have to compare it to so many other points Classifying new samples against the new reduced data set will sometimes lead to different results than comparing the new sample against the whole training set. This is the trade-o_ we have to make, between speed and accuracy. The CNN algorithm takes more time to run, especially on very huge data sets. There are optional methods for reducing dimensionality like an alternative version of CNN ie FCNN which often performs faster than CNN. Using CNN can cause our system to give us different classifications than if we just used kNN on the raw data. The FCNN algorithm initializes the consistent subset S with a seed element from each class label of the training set T. In particular the seed elements are employed which act as the centroids of the classes in T. The algorithm is incremental: during each iteration the set S is augmented until the stop condition is reached. The number of points contained in the subset could double at the end of each iteration. This is due to the fact that, for each point p such that Voren(p, S, T) where voren is a computing function, is not empty, a new point is added to the set S.

4. Conclusion

The concept of hubness can be used to cluster datasets containing high dimensional data. A lot of algorithms have been devised for clustering using the concept of hubness. The choice of the algorithm can be made depending upon the application for which we are going to use these algorithms. The fuzzy-based algorithms can work well when the data is either unbalanced or partially exposed as compared to other algorithms. The condensed nearest neighbour algorithm is best suitable for datasets which are order-independent. This is due to the fact that they make use of training datasets and approximate membership is found out which is done iteratively with each iteration providing better clustering. In this paper, it is shown that the different hubness algorithms can prove to be better than other types of clustering algorithms. The major advantage of the said methods is their efficiency in clustering, when high dimensional data which is order independent and partially exposed.

References

[1] Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, Mirjana Ivanovic."The Role of Hubness in Clustering High-Dimensional Data IEEE transactions on knowledge and data engineering, vol. 26, no. 3, march 2014.
[2] M. Radovanovi_c, A. Nanopoulos, and M. Ivanovi_c, "Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data," J. Machine Learning Research, vol. 11, pp. 2487-2531, 2010.
[3] N. Toma_sev, M. Radovanovi_c, D. Mladeni_c, and M. Ivanovi_c," Hubness -Based Fuzzy Measures for High-Dimensional k-Nearest Neighbor Classification," Proc. Seventh Int'l Conf. Machine Learning and Data Mining (MLDM), pp. 16-30, 2011.
[4] N. Toma_sev, M. Radovanovi_c, D. Mladeni_c, and M. Ivanovi_c, "The Role of Hubness in Clustering High Dimensional Data," Proc. 15th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD),Part I, pp. 183-195, 2011.
[5] N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "A Probabilistic Approach to Nearest-Neighbor Classification: Naïve Hubness Bayesian kNN," Proc. 20th ACM Int'l Conf. Information and Knowledge Management (CIKM), pp. 2173-2176, 2011.
[6] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer, "Local and Global Scaling Reduce Hubs in Space," J. Machine Learning Research, vol. 13, pp. 2871-2902, 2012.

Licensed Under Creative Commons Attribution CC BY

[7] N. Toma_sev and D. Mladeni_c, "Nearest Neighbor Voting in High Dimensional Data: Learning from Past Occurrences," Computer Science and Information Systems, vol. 9, no. 2, pp. 691-712, 2012.
[8] Sunita Jahirabadkar and Parag Kulkarni Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms,International Journal of Computer Applications (0975 – 8887) Volume 63– No.20, February 2013
[9] C. Ding and X. He, "K-Nearest-Neighbor Consistency in Data Clustering : Incorporating Local Information into Global Optimization,"Proc. ACM Symp. Applied Computing (SAC), pp. 584-589, 2004.
[10] I.S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-Means: Spectral Clustering and Normalized Cuts," Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 551-556, 2004.
[11] H. Bian, L. Mazlack, Fuzzy-rough nearest neighbor classification approach, in: Proceedings of the 22th International Conference of the North AmericanFuzzy Information Processing Society (NAFIPS'03), Chicago, Illinois, USA, July 24–26, pp. 500–505.

Licensed Under Creative Commons Attribution CC BY