STUDENT RESULT ANALYSIS USING CLUSTERING TECHNIQUE

[1]ALTAMAS KHAN & [2]DR. BONTHU KOTAIAH
[1]M-TECH, Dept. of CS & IT maulana azad national urdu university, gachibowli, Hyderabad, Telangana,
[2]Assistant professor, Dept. of CS & IT maulana azad national urdu university, gachibowli, Hyderabad, Telangana,

Abstract

Various clustering algorithms have been developed to crew data into clusters in diverse domains. Nevertheless, these clustering algorithms work with no trouble either on pure numeric data or on pure specific information, most of them perform poorly on combined categorical and numeric information forms. In this paper, a brand new two-step clustering approach is offered to find clusters on this type of data. On this procedure the objects in specific attributes are processed to construct the similarity or relationships among them founded on the ideas of co-incidence; then all express attributes may also be converted into numeric attributes founded on these built relationships. Subsequently, when you consider that all categorical information are modified into numeric, the present clustering algorithms can be applied to the dataset without anguish. Nonetheless, the prevailing clustering algorithms endure from some risks or weak point, the proposed two-step method integrates hierarchical and partitioning clustering algorithm with including attributes to cluster objects. This system defines the relationships amongst gadgets, and improves the weaknesses of applying single clustering algorithm. Experimental evidences exhibit that strong results will also be accomplished by way of applying this process to cluster blended numeric and categorical knowledge.

Keyword Words: clustering algorithms, educational data Set, K-means

1. INTRODUCTION

We handle clustering in virtually each facet of day-to-day life. Clustering is the area of energetic research in a number of fields corresponding to records, pattern realization, and desktop studying. In knowledge mining, clustering offers with very large information units with exclusive attributes associated with the information. This imposes designated computational specifications on relevant clustering algorithms. A type of algorithms has just lately emerged that meet these specifications and were efficiently applied to real lifestyles information mining issues [1]. Clustering methods are divided into two basic forms: hierarchical and flat clustering. Within every of these varieties there exists a wealth of subtypes and distinct algorithms for locating the clusters. Flat clustering algorithm intention is to create clusters that are coherent internally, and obviously unique from every other. The data inside a cluster must be as similar as feasible; knowledge in one cluster must be as assorted as

viable from documents in other clusters. Hierarchical clustering builds a cluster hierarchy that may be represented as a tree of clusters. Every cluster can also be represented as little one, a parent and a sibling to different clusters. Despite the fact that hierarchical clustering is superior to flat clustering in representing the clusters, it has a quandary of being computationally intensive in finding the vital hierarchies [8]. The preliminary goal of the venture is to make use of flat clustering methods to partition knowledge into semantically associated clusters. Extra, situated upon the clustering first-class and understanding of the info we increase cluster representation making use of hierarchical clustering. This may also effect in hybrid clusters between flat and hierarchical arrangement. Clustering algorithms furnished in the Apache Mahout library might be utilized in our work [2]. Mahout is a set of normally designed machine studying libraries. It is associated with Apache Hadoop [3] for large scale computer learning in allotted atmosphere. Presently Mahout supports in most cases suggestion mining, clustering and classification algorithms. For our mission we identified to evaluate a suite of clustering algorithms - okay-method, cover, fuzzy ok-manner, streaming k-manner, and spectral k-way to be had within the Mahout library. We've got used quite a lot of collections: web pages and tweet as our knowledge set to evaluate clustering. Given that clustering is an unmanaged classification finding the right number of clusters apriori to categorize the info is a elaborate predicament to deal with. The most effective approach to learn in regards to the quantity of clusters is to be taught from the info itself. We tackle this undertaking with the aid of estimating the quantity of clusters making use of ways like go-validation and semi-supervised finding out.

2. RELATED WORK

Clustering objects into groups is by and large established on a similarity metric between objects, with the intention that objects inside the same workforce are very similar, and objects between one-of-a-kind agencies are less equivalent. In this assessment we focus on record clustering for web sites and tweet information. The applying of textual content clustering can be each online or offline. Online functions are viewed to be extra effective in comparison with offline purposes in terms of cluster great, however, they endure from latency disorders. Text clustering algorithms may be labeled as flat clustering and hierarchical clustering. Within the next two subsections we tricky extra important points about these algorithms. 2.0.1 Flat clustering algorithms Flat clustering explains find out how to create a flat set of clusters with none explicit constitution that may relate clusters to each other [6]. Flat clustering ways are conceptually simple, however they have got a number of drawbacks. Most of the flat clustering algorithms, like k-manner, require a pre-unique quantity of clusters as enter and are non-deterministic. 2.0.2 Hierarchical clustering algorithms Hierarchical clustering builds a cluster hierarchy, or in different words, a tree of clusters. Suggests an instance of hierarchical clustering for a collection of points. Hierarchical clustering outputs are structured and extra informative than

flat clustering. Hierarchical clustering algorithms are additional subdivided into two types (1) agglomerative approaches - a bottom-up cluster hierarchy new release by means of fusing objects into corporations and corporations into greater clusters. (2) divisive methods - a high-down cluster hierarchy iteration by partitioning a single cluster encompassing all objects successively into finer clusters. Agglomerative approaches are extra most commonly used [10]. Hierarchical clustering does not require realizing the pre-detailed quantity of clusters. However this knowledge came with the cost of the algorithm complexity. Hierarchical clustering algorithms have a complexity that is at least quadratic in the quantity of records in comparison with the linear complexity of flat algorithms like ok-means or EM

3. Literature survey

A bigger schooling Predictive mannequin utilizing information Mining systems this paper lists a high scope for the students to make a decision for the brighter future with distinctive and correct evaluation. Because the effectively, accuracy, and effectiveness play the important position in the approach of Indian education approach, use of the Random wooded area procedure provides us an most fulfilling strategy to the real world scholar's education. On this paper, we've used the process of Random woodland to predict the profession decision for the twelfth passing out pupils. Using Random woodland has helped the students to take a correct appropriate resolution as per their interest and expertise. The final intention is to provide a greater perception to design a greater Indian education process for Indian scholars with the robust effect. This review could extend to greater aspects to resolve tricky choice databases in an effective method.

Information Mining in education Sector: A evaluate plenty of interest has been noticeable in EDM nowadays on the grounds that a tremendous quantity of pupils are enrolling for bigger schooling. Through EDM associations' researchers and stakeholders can bring more and more pleasure amongst scholar's fraternity. Academic information mining finds its utility no longer most effective in descriptive and predictive analytics but additionally in prescriptive analytics the place compatible moves can also be prescribed. Figuring out pupils, proper profiling and correct predictions is not going to simplest expand the first-class of schooling but in addition increase good finding out expertise to the scholars' fraternity. As a result of increasingly utilization of internet by means of scholars today, huge information is on hand about them. Via information mining, we are able to extract useful know-how that may support the schooling procedure to formulate right tactics for our youths

Discovery of Strongly associated subjects in the Undergraduate Syllabi making use of knowledge Mining knowledge mining contains a sort of methods that can be utilized to extract imperative and exciting talents from mammoth amounts of information. Data mining has been efficiently utilized in a type of domains to reap expertise gigantic in choice making. On this paper, we reward a real-world scan performed in an ICT educational institute in Sri Lanka. Our

experiment considers an information repository consisting pupils' efficiency in a giant ICT educational tuition. We apply a sequence of knowledge mining tasks to find relationships between topics in the undergraduate syllabi. This capabilities provides many insights into the syllabi of specific academic programmers and outcome in expertise important in selection making that instantly affects the best of the educational programmers.

4. CONCLUSION

Clustering has been broadly applied to quite a lot of domains to discover the hidden and priceless patterns within data. Due to the fact the most accumulated data in actual world contain both specific and numeric attributes, the ordinary clustering algorithm cannot manage this variety of knowledge readily. Hence, on this paper we endorse a brand new approach to explore the relationships amongst express objects and convert them into numeric values. Then, the prevailing distance centered clustering algorithms will also be employed to group mix types of knowledge. Furthermore, as a way to overcome the weaknesses of k-method clustering algorithm, a two-step approach integrating hierarchical and partitioning clustering algorithms is offered. The experimental outcome show that the proposed approach can acquire a high first-class of clustering results. On this paper, the TMCM algorithm integrates HAC and ok-manner clustering algorithms to cluster combined style of data. Applying other algorithms or sophisticated similarity measures into TMCM may just yield higher outcome. In addition, the quantity of subset i is set to 1-third of number of objects on this paper. Even though experimental results show that it is feasible, how to set this parameter precisely is valued at extra be taught sooner or later.

5. REFERENCES

[1] Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. “Data clustering: a review.” ACM Computing Surveys (CSUR) 31.3 (1999): 264-323.

[2] The Apache Mahout project’s goal is to build a scalable machine learning library. http://mahout.apache.org/

[3] Apache Hadoop. http://hadoop.apache.org/ Last accessed: 02/19/2015

[4] Wiederhold, G., Foreword. In: Fayyad U., Shapiro G. P., Smyth P., Uthurusamy R., editors, Advances in Knowledge Discovery in Databases. California: AAAI/ MIT Press, 1996;2.

[5] Han, J. and Kamber, K., Data mining: Concept and Techniques. San Francisco: Morgan Kaufman Publisher (2001).

[6] Jain, A. K. and Dubes, R. C., Algorithms for Clustering Data, New Jersey: Printice Hall (1988).

[7] Kaufman, L. and Rousseeuw, P. J., Finding Groups in Data: An Introduction to Cluster Analysis, New York: John Wiley & Sons (1990).

[8] Ng, R. and Han, J., Efficient and Effective Clustering Method for Spatial Data Mining, Proc. of the 20th

VLDB Conf. 1994 September. Santiago, Chile (1994).

[9] Zhang, T., Ramakrishman, R. and Livny, M., BIRCH:

an Efficient Data Clustering Method for Very Large Databases, Proc. 1996 ACM-SIGMOD Int. Conf. Management of Data, 1996 June. Montreal, Canada

(1996).