|| Volume 1 ||Issue3 ||June 2016||ISSN (Online) 2456-0774

INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH AND ENGINEERING TRENDS

REDUCTION THE DIMENSIONAL DEPENDENCE USING RANK-BASED SIMILARITY SEARCH

Miss. Manisha R. Ghunawat

P.G. Student, Computer science & Engineering, Everest Educational Society's Group of Institutions, Aurangabad, Maharashtra, India.

***Abstract*— The K-NN is a technique in which objects are classified depends on nearest training examples which is present in the feature query space. The K-NN is the simplest classification method in data mining. In K-NN objects are classified when there is no information about the distribution of the data objects is known. In K-NN performance of classification is depend on K and it can be determined by the choice of K as well as the distance metric of query. The performance of K-NN classification is largely affected by selection of K which is having a suitable neighbourhood size. It is a key issue for classification. This paper proposed a data structure which is for K-NN search, called as RANK COVER TREE to increase the computational cost of K-NN Search. In RCT pruning test involves the comparison of objects similar values relevant to query. In RCT, by assigning ranks to each objects and select objects with respect to their ranks which is relevant to the data query object. It provides much control on the whole execution costs of query. It gives experimental results which is a Non-metric pruning strategies for similarity search .when high dimensional data are used it gives the same result. It returns corrects query execution result in required time that relies on a intrinsic dimensionality of objects of the data set. RCT can exceed the performance of methods involving metric pruning and many selection tests involving distance values having numerical constraints on it**

Keywords:- K-Nearest neighbour search, intrinsic dimensionality, rank-based search, RCT.

I INTRODUCTION

In Data mining there is a various tools of data analysis which can find patterns of objects and relationships among the data. These tools make use valid prediction of object data. There are various fundamental operations such as cluster analysis, classification, regression, anomaly detection and similarity search. In all of which the most widely used method is of similarity search. Similarity search having built in principal of k-Nearest Neighbour (K-NN) classification. k-NN is founder of it. When number of data object classes is too large then similarity search produces low error rate as compare to other methods of analysis. Error rate of Nearest Neighbour classification shows when training set size increased is ‘asymptotically optimal ‘.In similarity search feature vectors of data objects attributes are modelled for which similarity measure is defined. Various application of data mining is depend on that. Similarity search can accesses an unacceptable –huge part of the data object elements, unless the other data can be distributed having special properties of data elements. Various Data mining application which uses common neighbourhood knowledge of data which is useful and having great meaning. High data dimensional tends to make this common information which very costly to gain. In Similarity search indices selection and identification of objects which is relevant to query objects depend on similarity values of information. This can measure the performance of similarity search. In distance-based similarity search make use of numerical constraints of similar values of data objects for building pruning and selection of data objects such types include the triangle inequality and additive distance bounds. The use numerical constraints shows large variations in the numbers of objects that can be examined in the execution of a query, It is difficult to manage the execution costs. To overcome the problem of large variation in objects analysis in execution. We build a new data structure, the Rank Cover Tree (RCT), used for k-NN. This can totally exclude the use of elements of data objects having numerical constraints. In RCT all internal selections operation are made using the ranks of that objects of data according to the query, having strict control of execution of data query. By using a rank of objects it gives rank-based search analysis provides best probability of analysis, the RCT gives a correct result of query in required time that fully depends on data set intrinsic dimensionality. The RCT is similarity search method use the ordinal pruning method and provides correct analysis of performance of the query result.

II LITERATURE SERVEY

For clustering, various effective and common methods require

the finding of neighbourhood sets of data objects which is depend on mostly at a required proportion of data set objects[1][2]. Various examples consists such as hierarchical (agglomerative) methods like ROCK [3] and CURE [4]; another method density-based example as DBSCAN [5], OPTICS [6], and SNN [7] and also non-agglomerative shared-neighbour clustering [8].

A recommender systems and anomaly detection technique used content based filtering approach [9], k-NN method also used in normal condition build, by making direct use of method k-NN cluster analysis. A another very popular local density-based measure that is method of local Outlier factor (LOF), totally rely on data set of k-NN whose computation to obtain the denseness of used data which is present in the test point of that section [10].

The application totally depend upon similarity search index can be improve its scalability and effectiveness. Researchers test practical techniques for speeding up the calculation of finding neighbourhood information at level of correctness. For application based on data mining tools, the methods consists of feature sampling which is used for used for local outlier detection method [11], for an k-NN classification method and its having own right of data sampling technique which is used in clustering as well as approximate similarity search method. BD-tree is an best suitable examples of fast approximate similarity search index method, a most-derivable standard for approximate k-NN search; it recognizable use of many rules of splitting and provides at early or near termination to obtain great performance approach of the KD-Tree. One of the most desirable technique for indexing is as Locality-Sensitive Hashing [12], [13], obtain best practical and formal performance of search for range queries methods by adjusting data object parameters that affect a exchange between time and accuracy. The most important technique that is for approximate search is spatial approximation sample hierarchy (SASH) similarity search method [14] had best outcome in speed up the performance of a algorithm shared-neighbour clustering [8], for a various data set object types.

III RANK COVER TREE

We proposed a new data structure which is a probabilistic used for similarity search index; the rank-based search means Rank Cover Tree (RCT), in which no involvement of numerical constraints for selection and pruning of data element objects. All internal operation such as selections of objects are made by consider to specified ranks of that objects element according to that query , having strict control on query execution costs. A rank-based probabilistic method having huge probability, the RCT perform a correct result of query execution in specific time that relies on a high portion of the intrinsic dimensionality of that data set.

Construction:

1. Consider each item **x** To X, provides **x** into levels 0,. . . x. Height of tree is h, x can follows technique of a geometric distribution with $q = jXj^{-1=h}$.

2. A partial RCT can be build by connecting each items in that level to an artificial root of tree on the highest level.

3. In partial RCT by using approximate nearest neighbors method which is found in the partial RCT can connect the next level of tree.

4. A RCT can be well-build with very high probability.

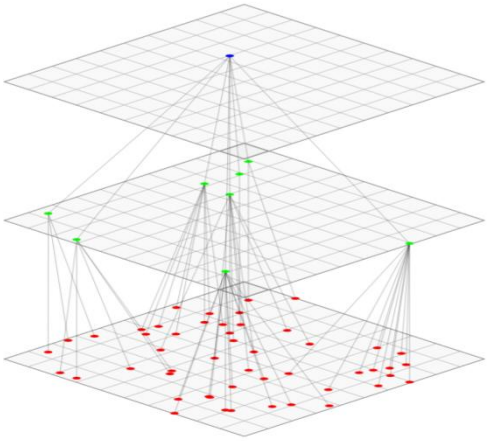

Figure 1 RCT Construction

To implement Rank Cover Tree it consists of design features of similarity search SASH and also design feature of Cover Tree. SASH can be used for approximate searching and cover tree for exact search of objects. both of these make use of a ordinal strategy for pruning of objects and it allows for strict control on query execution cost which is obtained with method of queries of approximate search . At each and every level of the tree structure visited the number of neighbouring nodes can be restricted, the user also reduces average required execution time of that query at the each level of that query accuracy. The proximity search of Tree-based strategies make use of distance metric method in two ways in which numerical constraint of objects among three data objects on its the distances as it is examined by the method of triangle inequality, or distance of data candidates from its a reference point of numerical (absolute) value constraint present on it.

I OBJECTIVE:

1. The RCT can increase the performance of methods that involves metric pruning strategy or other type of selection tests having numerical constraints on distance values.
2. To increase the computational cost of K-NN Search.
3. Using RCT user can minimize the average amount of time required for execution .to obtain a great query accuracy.

4. It provides tighter control on overall execution costs. Provides best result for similarity search

II. NECESSITY:

1. In RCT Rank thresholds method specifically calculate the number of data objects which is to be selected for pruning it avoid and reduce a major of variation of data elements objects in the overall execution time of query.
2. It improves computational cost of similarity search.

IV RELATED WORK

This paper consists of two most important and recently-developed approaches that are quite dissimilar from each other which is consider to proposed RCT data structure. The SASH heuristic is used for approximate searching of similarity, and second approach that is the cover Tree used for exact searching of similarity. RCT can use method of combinatorial search similarity approach. The SASH also used an combinatorial similarity search approach, whereas In the cover tree numerical constraints are used for selection and pruning of data objects. Description of SASH and Cover Tree as given below.

I. COVER TREE:

In Cover Tree the intrinsic dimensionality performance can be analyzed by a common search method for determining nearest neighbourhood data queries example. In this approach, a randomized structure can found like to be skip list which can be used to recognized pre-determined samples of data elements which is surrounding points object of interest. By performing same procedure by shifting the sample focus to sample data elements which is closest to the query object element and discovering new set of samples which is present in the surrounding points of interest. The rate of expansion of sample element S which is to be having minimum value of δ as required condition holds above, it is subjected to the various alternate of minimum ball element set size a.

The similarity search method Cover Tree gives a answer for exact similarity search of queries using an approach of early termination. In analysis of some algorithm basically focus is on expansion constant, it provides the results for exact nearest neighbour queries. Consider the S having the expansion constant C, following explicitly dependence on expansion rate C:

	Cover Tree	Nav. Net	[KR02]
Constr. Space	O(n)	C^o(1)n	C^o(1)n ln n
Constr. Time	O(c^6n ln n)	C^o(1)n ln n	C^o(1)n ln n
Insert/Remove	O(c^6 ln n)	C^o(1) ln n	C^o(1) ln n
Query	O(c^12 ln n)	C^o(1)) ln n	C^o(1) ln n

The algorithm which is work without knowing the information of that structure; and the analysis is done according to the some assumptions. In some algorithm (as in [KL04a] but not in [KR02]) work can be done without time complexity Comparison in terms of expansion constant C and it can be subtle .

II. SPATIAL APPROXIMATION SAMPLE HIERARCHY (SASH):

The huge amount of data sets objects that used a data structures providing the better performance for an amount of N data items within given database. The R-Tree plays an efficient role for efficiency of DBSCAN. To handle very massive data sets, use SASH technique. The SASH method can build minimal number of assumptions about associative objects queries metric. SASH does not regulate a partition of the query search space, as the instance of R-Trees can done. For similarity search of approximation of k-NN (k-ANN) queries present on the huge data sets, the similarity search SASH can systematically provide a huge part of k-NNs truth of queries at specific speeds of randomly of two different orders of relative size which is faster than regular sequential search method. For clustering method and navigation of very huge , very large

Dimensional text, image sets of data on which The SASH can perform successfully . The SASH internally can make use of a k-NN query on very small data sets elements. the SASH probably having a multi-level structure which is to be recursively constructed. In SASH building on random sample of half-sized data set $S' \subset S$ .the data object element set S, by connecting each and every object which is remaining present outside S' to the many of its nearest neighbourhood which is approximate of object from at an intervals S'. Many types of queries can be processed by its initial method by establishing approximate neighbours of objects at specific intervals of sample S', so that pre-established connections can be used to find various neighbours of objects at specific intervals of the information can be gain data set.

V RESEARCH WORK

We can overcome the drawback of operations involves an numerical constraints like the triangle inequality method or distance ranges in that data objects count actually proceed or examined having very high variation of objects, because of it the overall or complete time required for execution cannot be easily determined or predicted. Using RCT we can easily predict execution time. To increase the scalability and efficiency of data mining applications that fully rely on similarity search values. Finding the best methods for efficiently speed up the computational power of nearest neighbourhood information at the great expense of accuracy.

VI CONCLUSION

We have developed rank-based search that is Rank Cover Tree which is a new data structure for k-NN search. In which ordinal pruning approach is used that involves direct distance values of data objects comparisons. The RCT construction is independent on the representational high dimension of the data. but it can be probabilistically analyzed in the form of approach a measure values of intrinsic dimensionality. The RCT can be build by using two main methods –that means the cover Tree and SASH structures techniques.

ACKNOWLEDGEMENT

It is my great pleasure in expressing sincere and deep gratitude towards my guide Prof.Yogesh R.Nagargoje. I am also thankful to Head of Department of Computer Science and Engineering, Prof. Rajesh. A. Auti for providing me various resources and infrastructure facilities. I also offer my most sincere thanks to Principal of Everest College of Engineering, Aurangabad, my colleagues and staff members of computer science and Engineering department, Everest college of Engineering, Aurangabad for cooperation provided by them in many ways.

REFERENCES

[1] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, San Francisco, CA, USA: Morgan Kaufmann, 2006. [1] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, San Francisco, CA, USA: Morgan Kaufmann, 2006.

[2] T. Cover, and P. Hart, "Nearest neighbor pattern classification,"IEEE Trans. Inf. Theory, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[3] S. Guha, R. Rastogi, and K. Shim, "ROCK: A robust clustering algorithm for categorical attributes," Inf. Syst., vol. 25, no. 5,pp. 345–366, 2000.

[4] S. Guha, R. Rastogi, and K. Shim, "CURE: An efficient clustering algorithm for large databases," Inf. Syst., vol. 26, no. 1, pp. 35–58,2001.

[5] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in Proc. 2nd Int. Conf. Knowl. Discov. Data Mining, 1996,pp. 226–231.

[6] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS:Ordering points to identify the clustering structure," in Proc. ACM SIGMOD Int. Conf. Manag. Data, 1999, pp. 49–60.

[7] L. Ert€oz, M. Steinbach, and V. Kumar, "Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data," in Proc. 3rd SIAM Int. Conf. Data Mining, 2003, p. 1.

[8] M. E. Houle, "The relevant set correlation model for data clustering," Statist. Anal. Data Mining, vol. 1, no. 3, pp. 157–176,2008.

[9] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, pp. 1–58, 2009.

[10] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "LOF: Identifying density-based local outliers," SIGMOD Rec., vol. 29, no. 2,pp. 93–104, 2000.

[11] T. de Vries, S. Chawla, and M. E. Houle, "Finding local anomalies in very high dimensional space," in Proc. IEEE Int. Conf. Data Mining,2010, pp. 128–137.

[12] A. Gionis, P. Indyk, and R. Motwani, "Similarity search in high dimensions via hashing," in Proc. 25th Int. Conf. Very Large Data Bases, 1999, pp. 518–529.

[13] P. Indyk, and R. Motwani, "Approximate nearest neighbors: Towards removing the curse of dimensionality," in Proc. 30th ACM Symp. Theory Comput., 1998, pp. 604–613.

[14] M. E. Houle and J. Sakuma, "Fast approximate similarity search in extremely high-dimensional data sets," in Proc. 21st Intern. Conf. Data Eng., 2005, pp. 619–630.

[15] M. E. Houle and M. Nett, "Rank cover trees for nearest neighbor search," in Proc. Int. Conf. Similarity Search Appl., 2013,pp. 16–29.