International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(1), pp.030-036
DOI: http://dx.doi.org/10.21172/1.81.004
e-ISSN:2278-621X

SPATIAL DATA MINING FOR FINDING NEAREST NEIGHBOR AND OUTLIER DETECTION

Srishty Jindal[1] and Dr. Kamlesh Sharma[2]

Abstract- Spatial data mining is a process to extract interesting patterns related to space. Space can be geographic space, the universe, a VLSI design, a molecular structure, or a human body. With the proliferation in use of spatial databases the probability of getting outliers is also increased. These outliers can be noisy data or highly valuable information. If the noise exists in the database, the performance of data mining algorithm may be degraded [14]. Detection of outliers in spatial database can be area of research in various applications. Spatial databases can be used in location based services (e.g. Google maps) to find nearest neighbors. If there is a data point which is not nearer to other data points, then this data point is considered as outlier. In this paper, we have discussed various researches for finding nearest neighbor and outlier detection.

Keywords- Data Mining, Spatial Databases, Outliers

I. SPATIAL DATABASES

A spatial database is a collection of records related to space. The space can be a geographic space, a human body or a VLSI chip [15]. In SDBMS, objects are defined in a geometrical shape such as points, lines and polygons. Spatial database offers spatial data types, data models and query languages to process the spatial data. The major components of spatial database include a data model, query languages, processing and optimization tools, and indices. Examples of spatial databases include weather climate data, river, farms, medical imaging etc. Spatial databases support spatial indexing, efficient algorithms for processing spatial operations, and domain specific rules for query optimization. Spatial databases can be used in medical, astrology, biology, and defense and in many more areas.

A geographic information system (GIS) or geospatial information system is designed to capture, store, manipulate, analyze, manage, and present all types of spatial or geographical data. In GIS there can be many problems that include finding the nearest neighbor or finding the outliers in a data set.

[1] *Research Scholar, School of Computing Lingaya's University, Faridabad, Haryana, India*
[2] *Associate Professor, School of Computing Lingaya's University, Faridabad, Haryana, India*

II. DATA MINING

Data mining [1] is a process to extract, analyze the data from many perspectives and transform it into an understandable and useful structure for further use. The basic goal of data mining is to extract the patterns and knowledge from large amounts of data, not the extraction (mining) of data itself. This can be used for further prediction in many areas like weather forecasting, business, fashion and textile industries etc.

Data mining is used to extract interesting patterns (cluster analysis), unusual records (anomaly/ outlier detection) and dependencies (mining association rules). These patterns can be used in machine learning and predictive analysis. Data mining involves following important tasks:

i) **Anomaly/outlier detection:** Detection of unusual data records that may not lie in any of the cluster and might be interesting data the data requires further investigation.
ii) **Association rule:** Finds the relationship between variables. For example, bakery might gather data on purchasing habits of customers. Using association rule learning, the bakery determines which item sets are frequently bought by customers and then use this information for increasing the revenue. This is referred as market basket analysis.
iii) **Clustering:** is the process of discovering groups/clusters in the dataset that are similar. These clusters are not known previously and are a type of unsupervised learning.
iv) **Classification:** is the process of generating known structure in a data set. Classifiers are constructed depending on the training data to classify. These classifiers are then further used for classification of records.
v) **Regression:** attempts to find a function which models the data with the least error.
vi) **Summarization:** is the process of reducing the size of data without changing its actual meaning.

Data analysis [2] can be done using Artificial Neural Network, Genetic Algorithm, Decision trees, Nearest Neighbor, and Rule Induction. Neural Network is based on biological neuron structures that learn through training. Genetic Algorithms involves the genetic combination of two individuals to provide optimized result. GA includes process of selection, combination and mutation based on concept evolution. A tree based structure, decision tree, is used for generating rules that can help in classification of data objects. In this tree, each branch represents a set of decision. NN classifies each record in a data set most similar to it in historical dataset. Based on statistical data, useful rules can be extracted by rule induction. These rules can be represented in if-then-else-form.

III. OUTLIER DETECTION

An outlier is any person or thing which is inconsistent, different from other members in a group. Outlier can have many anomalous causes. It may cause due to experimental error, instrumental error, human error or a deviation in population. Outlier can be a defect in assumed theory which

can be rectified by further investigation. Although outliers are considered as noise or error in dataset, still they may carry useful information. Hence, it is essential to find outliers before analysis; otherwise it may lead to incorrect results. In spatial databases, the detection of outlier is the method of identifying spatial objects having distinct features from its surrounding objects. Detection of spatial outliers helps in getting useful information from large database.

In section IV we discuss the various algorithms used by different researchers for clustering and classification. These two techniques can further be useful in finding nearest neighbor and detecting outliers. Section V describes various application areas where these algorithms can be applied for the purpose of getting task easier. And finally, in section VI, conclusion and future scope is given.

IV. RELATED WORK

Derya Birant [3] proposed an algorithm to find inconsistent outliers in large databases. Primarily, clustering is done to find the outliers. For clustering, an improved Density-based spatial clustering of applications with noise algorithm is used to enhance the proficiency of algorithm. Outliers with a little difference in their neighborhood locations are not considered as outliers. DBSCAN algorithm discovers the clusters of arbitrary shape and good efficiency in large database. After clustering, potential outliers are checked to verify whether these objects are actually spatial outliers(S-outliers) or not. For verification background knowledge is required. If background knowledge is not available, then neural networks can used to handle the problem. These S-outliers are then observed in consecutive time units. S-outliers are compared with other objects at same level in different time intervals. Then ST outliers are identified. Improved R-tree is used for implementing the same. Spatial object were represented by nodes. These nodes were linked in temporal order. DBSCAN algorithm for clustering and Local Outlier factor for detecting outliers can be applied together to increase the feasibility of framework [19].

Amitava Karmakar [4] introduced an partitioning based clustering algorithm for detecting errors in spatial databases and minimizing outliers. In this technique, data set D is partitioned into K partitions. Then n local clusters are generated in each partition of data set. Each cluster has one cluster center. Apply clustering algorithms to all the cluster centers and generates a single cluster in the data set. Dissimilarity between the cluster center of final cluster and center cluster is calculated. This difference indicates how correct the database is.

Michael minoch [5] implemented an distance based clustering algorithm to handle the vague terms used in spatial database queries. Vague terms such as ‘near’ and ‘far’ create a difficulty in mapping natural Language interface and database. These terms can be True/False at the same time. So their value depends on the context of the query. All the vague terms are context dependent. To handle the problem of vague terms supervaluation technique is used. Values which are true in every context are termed as super true and the values which are false in every

context are super false. A threshold value is used to find super true and super false values. For the rest of the values a new parameter should be introduced each time to understand the context and clear the vagueness. It is tricky to find the point when a new parameter should be introduced.

Cha LunLi [6]introduced a bi-chromatic reverse k-nearest neighbor algorithm for top-n query processing in spatial databases. This algorithm can be applied on two different types of data sets. To efficiently answer the query, a Voronoi diagram is constructed. This diagram is used to rank the data points nearest to the query point. RkNN can be computed by Voronoi tree which is composed of voronoi diagram and R-tree. With the help of the information associated with voronoi diagrams, the upper bound of cardinalities of answer set of BRkNN query can be computed quickly. According to the ranking of answer set, top n query can be answered efficiently. Based on Voronoi diagram and an existing approach to answer RkNN queries, they proposed a method to find candidate region which shortens the search space. According to INCH algorithm, data plane is divided in two half planes using bisector. If the query point lies in half plane then searching nearest neighbor in q half plane can be avoided.

M. H. Marghny [7]proposed an outlier detection algorithm using Improved Genetic K-means algorithm. This algorithm includes two stages: In first stage, improved genetic k-means algorithm is used for clustering and in second stages all the vectors which are far from their cluster centroids are removed iteratively. Outliers can be detected using Indegree number and by outlier removal clustering, distance based approach. In Indegree outlier detection, a kNN graph is drawn in which each data vector is a vertex and edges are pointers to its neighbors. Weight of edge is calculated by finding the distance between data vectors. The outlyingness factor is defined for each data vector. If the outlyingness factor is above threshold value, then it is considered as outlier. In outlier removal clustering, outliers can be detected using the distance between data vector and cluster centroid. In each iteration, data vector with maximum distance is found. It works well on randomized data. The proposed algorithm is similar to ORC and using IGK-means algorithm for clustering.

Karanjit singh [8] highlighted some outlier detection techniques and its applications in various research areas. Complexity of outliers in various domains is also explained. Depending on the nature, outliers can be classified in three categories i.e. point, contextual and collective outliers. Point outliers are simplest and most common topic for research. If an instance is abnormal in respect of other data instances, then it is a point outlier. Credit card fraud detection is an example of point outlier because fraud can be detecting with the help of one feature, amount spent. Contextual outliers are the one which are abnormal in respect of other data instances in some context, otherwise not. For example if a person is 6 feet tall then it is normal but if a kid is 6 feet tall then it is considered as contextual outlier. Collective outlier is a collection of related outliers. Individual data instance in this group may not be an outlier. For example, In ECG output, high, low and nominal values may be normal. But any constant value for a long time may be an outlier. Collective outliers can be found in sequential, graph and spatial data. After detecting

outliers, a score and label must be defined to find the degree of extent to which that instance is an outlier.

Ke Deng [9] introduced Best Keyword Cover Search Technique to find the optimized solution for a nearest neighbor query using R-tree. mCloset Keyword cover(mCK)[12,11,10] can be used to find objects which includes all query keywords and have minimum interobject distance. The proposed keyword-NNE algorithm gives results according to all the query keywords, minimum interobject distance, and keyword ratings. Their baseline algorithm returns a large number of candidate solutions. In this case performance of the algorithm may drop significantly when more query keywords are given. Keyword-NNE algorithm search for local best solution and process each keyword candidate cover such that there are very less candidate covers. Many web services provide facilities to the users to put text constraints on geographical location. These types of problems needs more focus on textual data in spatial queries [17].

Ian De Felipe [13] proposed an algorithm using R-tree used to find the objects closer to query location and that contains a set of keywords also. These keywords include the attributes of the objects. In spatial databases, if any user initiates a nearest neighbor query with some specific attributes, its baseline algorithm first find the objects having all these attributes by intersecting the keywords. It will only find the objects having all the keywords. After finding the objects, their interobject distance is calculated. Then top k results are obtained according to the distance. For example, if a user wants to find a hotel with wireless internetworks and swimming pool, first it will list all the hotels with these two amenities, after listing such hotels, it will find the interobject distance. According to the distance, it will return the solution. The only drawback with this technique is if there will not be any object having all the keywords, then it will backtrack its search and again look for solution with less keywords. A spatial keyword query takes the user's current location and query keywords as input arguments and returns the objects which are spatially and textually related to the input arguments [16].The combination of geo-location and text document enables a top-k query which considers both location and text proximity [18].

In some cases outlier detection is considered similar to the clustering and classification problem. The major troubles with nearest neighbor problem can be solved using clustering or classification. Several clustering algorithms can be used to group the data points. After grouping the data, there are some data points which are not the part of any group. These residues are termed as noisy data or irrelevant data. This noisy data can be handled in many ways.

V. APPLICATIONS

Any disturbance in ecosystem can be a part of outlier detection. Outliers may be found in Meteorological data that include all kinds of disasters such as, cyclone, earthquake, floods etc. [20]. Outlier detection plays vital roles in astronomy for finding any abnormal activity or for

weather prediction also. Outliers can be found in biological data for finding any type of medical change in the body. Analyzing census data gives the growth rate in population which can also be helpful in predicting the outliers. Traffic volume can also analyzed such as during National Holidays there are high chances of people moving from one place to another. Outliers can also be used for Crime mapping. In Public health analysis, outliers can detect occurrence of a particular disease in a particular area. For example tetanus, spread over many hospitals in a city indicates the trouble in vaccination program in that city.

VI. CONCLUSION AND FUTURE SCOPE

This paper first introduce about the data mining in spatial databases. Spatial data mining can be done for detecting any abnormal activity in spatial database. This abnormal activity can be helpful to the people in any way or it may cause harm. So detection of these activities is a recent topic for research. Clustering/ classification algorithm helps in finding the outliers. In our further research we will try to work upon the improving performance in terms of speed and cost factor.

REFERENCES

[1] Manjula Aakunuri, Dr. G. Narasimha, Sudhakar Katherapaka,"Spatial Data Mining: A Recent Survey and New Discussions", International Journal of Computer Science and Information Technologies (0975-9646),Vol 2(4), 2011, pp 1501-1504.

[2] Divya Goyal, Hardeep Singh,"Survey paper on Data Mining Techniques: Outlier Detection and Text Summarization"International Journal oScientific & Engineering Research(ISSN 2229-5518), Volume 5, Issue 3, March 2014.

[3] Derya Birant, Alp Kut,"Spatio-Temporal Outlier Detection in Large Databases", Journal of Computing and Information Technology - CIT 14, 2006, 4, 291–297.

[4] AmitavaKarmaker, Syed M.Rahman,"Outlier Detection in Spatial Databases Using Clustering Data mining", sixth international Conference on Information Technology: New Generations , 1657-1658, 2009.

[5] Michael Minock, Johan Mollevik,"Context-dependent 'near' and 'far' in spatial databases via supervaluation" Data and Knowledge Engineering, Elsevier , Vol 86, July 2013, Pages-295-305.

[6] Cha-LunLi, En Tzu Wang, Guo-Jhu Huang, Arbee L.P. Chen," Top-n query processing in spatial databases considering bi-chromatic reverse k-nearest neighbors" Information Systems, Elsevier, 2014, pages 123-138.

[7] M.H. Marghny, Ahmed I.Taloba,"Outlier Detection using Improved Genetic K-means", International Journal of Computer Applications(0975-8887), Vol 28-No.11, August 2011.

[8] Karanjit Singh, Dr. SuchitaUpadhyaya," Outlier Detection: Applications and Techniques", International Journal of Computer Science Issues, Vol 9, Issue 1, No3, January 2012, pages 307-323.

[9] Ke Deng, Xin Li, Jiahenglu, XiaofangZhou,"Best Keyword Cover Search", IEEE Transactions on knowledge and Data Engineering, Vol 27, no.1, January 2015, 61-73.

[10] D. Zhang, B.Ooi ,A.Tung,"Locating mapped resources in weg 2.0", in Proc.IEEE 26th Int. conf. Data Eng.,2010,pp.521-532.

[11] D. Zhang, Y. Chee, A. Mondal, A. Tung, M. Kitsuregawa,"keyword search in spatial databases: Towards searching by document", in Proc. IEEE Int. Conf. Data Eng,. 2009,pp 688-699.

[12] X. Cao, G. Cong.,and C. Jensen,"Retrieving top k-prestige based relevant spatial web objects" Proc. VLDB Endowment, Vol 3, nos.1/2, pp 373-384,Sep2010.

[13] Ian De Felipe, vagelisHristidis, Naphtali Rishe," Keyword Search on Spatial Databases", In Pro.IEEE 24th Int. Conf. on Data Engineering, pages 656-665. April 2008.

[14] Yushan Qiu, Xiaoqing Cheng, Wenpin Hou, and Wai-ki Ching," On Classification of Biological Data using Outlier Detection", 12th International symposium on operations research and its applications in Engineering, Technology and Management (ISORA, 2015)page 140-150., aug 2015.

[15] Shashi shekhar, Siva Ravada, Xuan Liu, "Spatial databases-Accomplishments and Research Needs", IEEE transactions on Knowledge and Data Engineering, Vol 11, No. 1, January 1999.

[16] Xin Cao, Lisi Chen, Gao Cong, Christian S. Jensen, Qiang Qu," Spatial Keyword Querying" Springer-Verlag Berlin Heidelberg 2012, pp 16-29, 2012.

[17] Maria Christoforaki, Jinru He, constatinos Dimopoulos, Alexander Markowetz, Torsten Suel," Text vs space: efficient geo-search query processing" In CIKM, pp 423-432, October 2011.

[18] Gao Cong, Christian S. Jensen, Dingming Wu," Efficient retrieval of the top-k most relevant spatial web objects" VLDB Endowment, pp 337-348, August 2009.

[19] Mrs. Neeta M. Dumble, Mr. Bharat Tidke," A Framework for Outlier Detection In Geographic Spatial Data", International Journal in Foundations of Computer Science and Technology, Vol 5, No.2, March 2015.

[20] Jiang Zhao, Chang-Tien Lu, Yufeng Kou," Detecting Region Outliers in Meteorological Data", GIS'03, November2003.