IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 37-42
www.iosrjournals.org

K-NN Process In Applications Of GIS Technologies Using Spatial Temporal Data

Gandhimathi.D[1], Kanmani.K[2], Kirupa.S[3],
[1]*Assistant Professor, Csa&Ss, Sri Krishna Arts And Science College, Coimbatore*
[2]*Sri Krishna Arts And Science College, Coimbatore*
[3]*Sri Krishna Arts And Science College, Coimbatore*

Abstract: *In modern world user have started to search their requirement through their mobile User can access information through Mobile environment more easily regardless of user location. Accessing information can be of any type of searching a location. Searching can be to identify their nearest educational Institution, Petrol Bunk, Play area, Restaurant and so on. Spatial queries as utilize to access better information through the mobile at point of the world. KNN, Range query are popularly available to identify the required location.KNN can also apply for business purpose such as Banking, statistical process, Social network. In banking sector KNN is utilized to identify the nearest ATM for transaction, Loan decision, Bank credit risk analysis with k-nearest neighbor classifier. The K-NN is modified as K-NN classifier for the above process under banking sector. K-NN can be modified as K-NN Extract for Business application to identify the Product analysis through its sales, Predict the consumer behavior, Locating the retail & service for sold products. Classifier can be a simple matter of locating the nearest neighbor in instance space and labeling the unknown instance with the same class label as that of the located (known) neighbor. This approach is often referred to as a nearest neighbor classifier. The high degree of local sensitivity makes nearest neighbor classifiers highly susceptible to noise in the training data.*

Robust statistical methods are developed for many common problems, such as estimating location, scale and regression parameters. The Robust models can be achieved by locating k, where k > 1, neighbors and letting the majority vote decide the outcome of the class labeling. A higher value of k results in a smoother, less locally sensitive, function. The nearest neighbor classifier can be regarded as a special case of the more general k-nearest neighbor's classifier. The drawback of increasing the value of k is of course that as k approaches n, where n is the size of the instance base, the performance of the classifier will lead to K-NN Extract, to extract the required statistical baseline, the assumption that all unknown instances belong to the class most frequently represented in the training data.

Keywords: *K-NN, spatial, queries, classifiers*

I. Introduction

Data mining is processed by companies to turn raw data into useful information. Data mining (DM) is a part of KDD, an essential process to uncover the hidden information for evaluation. The principle of data mining can apply to various kind of data base such as relational, object-oriented ,Special applications- i.e., Multidimensional data base, Spatial Data base, temporal data base. An application of DM techniques is spatial data mining (SPDM). The process of extracting spatial relation and store in spatial database relate to SPDM. Mining of spatial data lead to identifying spatial data base consist of spatial related information are represented in multidimensional data with explicit knowledge object , position in space. Spatiotemporal data mining is an impending research area committed to various analysis of large Spatial temporal databases. Spatial Temporal Data consist of spatial properties as location of geometry of an object. Temporal property are time interface for which an object are valid .Various task of spatial temporal DM are multidimensional analysis of a spatial temporal data are time location, geometry are stored in SDW.[1].Spatial temporal neighborhood of a task of multidimensional data.Neighborhood of an object are used for Spatial temporal neighbors. The time dependent geometries are moving object is an applications of Geographical Information System (GIS).The moving object i.e., Vehicle, cell phone etc are embedded with GPS devices. The database of an tracking object are available through GPS device with the help of GIS techniques. The knowledge about location are identified and can frame a cluster to identify the nearest required data for an moving object. This paper is organized to start discussion with basics of Spatial Data Mining , its techniques & it is purpose .It is followed with Spatial temporal Data , GIS Technologies. The content is proceeded with Algorithms of SPDM, Continue with the discussion of KNN with the Banking data set. An Analysis on KNN algorithm utilization in weka tools by means of Synthetic data set is processed

II. Spatial Data Mining

Spatial data mining is the application of data mining in which the analysts use geographical and spatial information to produce business intelligence that require specific techniques and resources to get the geographical data into relevant and useful formats. Challenges involved in spatial data mining are identifying patterns and finding objects. It is used to find useful and non-trivial patterns in data. The goal of spatial data mining is to distinguish the information to build real, randomized spatial modeling and irrelevant results [a]. Spatial data mining differs from data mining through its attributes of neighbors of selected object .The explicit location with the extension of spatial objects define implicit relations of spatial neighborhood object which are utilized for Spatial data mining process. Fayyad et al. 1996, Miller and Han 2001 classifies Spatial data mining tasks and techniques into five categories including segmentation, dependency analysis, deviation and outlier analysis, trend discovery, and generalization and characterization.

2.1 Techniques in spatial data mining:

Since spatial data mining is an interdisciplinary subject, there are various techniques that include probability theory, evidence theory, spatial statistics, cloud model, neural network, genetic algorithms and decision tree, etc,. The implementation of spatial data mining is still needed to be studied[b].).It is a process of extracting spatial relation and store in spatial database relate to SPDM[c].Mining of Spatial data related information are requested as Multidimensional data with explicit knowledge object ,position in space

2.2 Purpose of spatial data mining:

The purpose of data mining is combining the structured and unstructured data to provide the workers to use it easily [b].

The steps in which the analysis is divided are:

- ✓ Deciding on business objectives- to understand the issues in the business objectives to create a suitable data mining model.
- ✓ Accessing the current situation the activity involves inventory of resources, requirements, constrains, risks and accidents etc.
- ✓ Deciding the data mining process goal – it is necessary to interpret the users Requirements.
- ✓ Creating the project plan – once the goals are set, it is possible to create a data mining project [b].

2.3 Algorithm used in spatial data mining:

KNN is one of the algorithms that are very simple to understand but works well in practice. It is also called as non parametric lazy learning algorithm. This means it does not use the training data points to do any generalization. There is no explicit training phase or it is very minimal. The training phase is pretty fast. Lack of generalization means the KNN keeps all the training data and all the training data sets are needed during the testing phase [c].

III. Spatial Temporal Data Mining

Spatial temporal knowledge discovery are a part of Geographic knowledge discover(GKD). Geographic data mining are closely related to the field of spatial databases, knowledge discovery and data mining. Spatial Data base with its related techniques can apply to various applications of GKD .Spatial and temporal dimensions add extensive intricacy of data mining process .Representation of Spatiotemporal data are continuous of spatial data, It also authority of collocated neighboring Spatial temporal objects. Analysis of large volume of spatiotemporal data without fixing any dimension is very difficult and complex .Spatial temporal analysis are categorized as temporal data analysis, spatial data analysis ,static spatiotemporal data analysis, dynamic Spatial temporal data analysis. Using fixed spatial dimensions and analyze how thematic attributes are change with time are temporal data analysis. Example are : Analysis of temperature, rainfall with respective over a period of time . Analysis of how thematic attributes data changing with respect to a distance from a spatial reference at a specified time. Example: Static spatiotemporal data analysis is studies the spatial dimensions for fixed temporal and thematic attributes dimensions. It can be applied to finding location which have rainfall in a particular time._ Dynamic Spatial temporal data analyses are defined as analyzing thematic attributes, spatial properties with respective to time. Applications domains of spatiotemporal dataset are: Medicine, Biology, Crop, Forestry, Ecology, Geographical, Meteorology, transportation, remote sensing, satellite telemetry, monitoring environmental resources and geographic information systems (GISs).

The reason to go for Spatial temporal is Spatial information are stored as Pixel values in Memory. Temporal image are Video of image frame sequence. With respect to time the frames are changed in video as temporal information. Spatial-temporal objects consist of finitely many spatial slices (i.e., time intervals), coordinates are linear functions of time on each slice, segments may degenerate but cannot rotate within a slice, Spatial-temporal object are polyhedron. Spatial statistics concerned with the analysis and modeling spatial data[a].

3.1. Spatial temporal data

It is used to manage both space and time information. It form a major amount of data generated by the GIS technologies, mobile devices,vision applications etc…., it have been used to extract the data from an database or file

Spatial-temporal data are abundant, and easily obtained. Examples are satellite images of parts of the earth, temperature readings for a number of nearby stations, election results for voting districts and a number of consecutive elections, trajectories for people or animals possibly with additional sensor readings, disease outbreaks or volcano eruptions.[6] Various spatiotemporal data mining task are: Multidimensional analysis of spatiotemporal data, Spatiotemporal characterization, spatiotemporal Topological Relationship discovery, Mining spatiotemporal Topological Relationship Patterns, Spatiotemporal Neighborhood, Spatiotemporal Association Rules, Spatiotemporal data classifications, Trend prediction or Detection, Spatiotemporal Data clustering, Spatiotemporal outliner analysis, Spatiotemporal collocation pattern or episode discovery, Discovering Movement patterns, Cascading spatiotemporal pattern discovery.[e]

IV. GIS Technologies

A system of hardware, software, and procedures designed to support the

- capture,
- management,
- manipulation,
- analysis,
- modeling and
- display of spatially-referenced data (located on the earth surface) for solving complex planning and management problems.”[4]

GIS store and manage huge volume of geographical entities such as road sections or lakes. Each entity consist of points, rectangle, surface of higher dimensional data location description of satellite image. The attributes of spatial objects are related to its neighbors of the object which is to extract .Specific location attributes are partitioning the database based on the identified locations use as a component of spatial data. Which are extracted through GIS techniques .Spatial data mining methods make intensive use of spatial relationships which is stored in spatial data warehouse.[d]

4.1 Purpose of GIS:

It allows the geographic features in real world locations to be digitally represented and stored in a database so that they can be abstractly presented in map form, and can also be worked with and manipulated to address some problem.[4] A GIS is an analytical tool used to integrate tabular information and graphical information. A geographic component can be identified in business data by an address or zip code and mapped accordingly. GIS is responsible for extracting for spatial data in multidimensional model. Spatial temporal Data consist of spatial properties as location of geometry of an object.

V. Assumptions In KNN

KNN assumes that the data is in a feature space .Each of the training data consists of a set of vectors and class label associated with each vector. It will be either + or – (positive or negative classes). But KNN can work equally with arbitrary number of classes [c].

We are also given a single number “k”. The number decides how many neighbors influence the classification. It is usually a odd number if the number of classes is 2. If k=1, then it is simply said as nearest neighbor algorithm[c].

5.1KNN for density estimation:

The classification remains the primary application of KNN, we can use it to do density estimation also. Since the KNN is non parametric, it can do estimation for arbitrary distributions. This idea is similar to PARZEN WINDOW. For estimating the density at point X, place the hypercube at the center of X and keep increasing the size of k till the neighbors are captured. Now estimate the density using the formula,

p(x) = {k/n}/{V}

Where n is the total number and V is the volume of hypercube[c].

5.2KNN for classification:

Case 1: k=1 or nearest neighbor rule

Let x be the point to be labeled. Find the point closest to x. Let it be y. Now nearest neighbor rule asks to assign the label of y to x. This seems too simplistic and sometimes even counter intuitive. If you feel that this procedure will result a huge error, you are right – but there is a catch. This reasoning holds only when the number of data points is not very large [c].

If the number of data points is very large, then there is a very high chance that label of x and y are same. An example might help – Let's say you have a (potentially) biased coin. You toss it for 1 million time and you have got head 900,000 times. Then most likely your next call will be head. We can use a similar argument here [c].

An informal argument is - Assume all points are in a D dimensional plane. The number of points is reasonably large. This means that the density of the plane at any point is fairly high. In other words, within any subspace there is adequate number of points. Consider a point x in the subspace which also has a lot of neighbors [2]. Now let y be the nearest neighbor. If x and y are sufficiently close, then we can assume that probability that x and y belong to same class is fairly same – Then by decision theory, x and y have the same class [c].

The book "Pattern Classification" by Duda and Hart has an excellent discussion about this Nearest Neighbor rule. One of their striking results is to obtain a fairly tight error bound to the Nearest Neighbor rule. The bound is P*<= P <= P^* (2 - {c}/{c-1} P^*)

Where P* is the Baye's error rate, c is the number of classes and P is the error rate of Nearest Neighbor. The result is indeed very striking (atleast to me) because it says that if the number of points is fairly large then the error rate of Nearest Neighbor is less that twice the Baye's error rate. Pretty cool for a simple algorithm like KNN[c].

Case 2: K or k-Nearest neighbor rule:

This is a straightforward extension of 1NN. Basically what we do is that we try to find the k nearest neighbor and do a majority voting. Typically k is odd when the number of classes is 2. Let's say k = 5 and there are 3 instances of C1 and 2 instances of C2. In this case, KNN says that new point has to labeled as C1 as it forms the majority. We follow a similar argument when there are multiple classes [c].

One of the straight forward extension is not to give 1 vote to all the neighbors. A very common thing to do is *weighted KNN* where each point has a weight which is typically calculated using its distance. For ex. under inverse distance weighting, each point has a weight equal to the inverse of its distance to the point to be classified. This means that neighboring points have a higher vote than the farther points [c].

It is quite obvious that the accuracy might increase when you increase k but the computation cost also increases [c].

5.3Applications of KNN

1.Nearest Neighbor based Content Retrieval

This is one the fascinating applications of KNN – Basically we can use it in Computer Vision for many cases – You can consider handwriting detection as a rudimentary nearest neighbor problem. The problem becomes more fascinating if the content is a video – given a video find the video closest to the query from the database – Although this looks abstract, it has lot of practical applications – Ex: Consider **ASL** (American Sign Language) . Here the communication is done using hand gestures [d].

So lets say if we want to prepare a dictionary for ASL so that user can query it doing a gesture. Now the problem reduces to find the (possibly k) closest gesture(s) stored in the database and show to user. In its heart it is nothing but a KNN problem[d].

2.GeneExpression

This is another cool area where many a time, KNN performs better than other state of the art techniques. In fact a combination of KNN-SVM is one of the most popular techniques there. This is a huge topic on its own and hence I will refrain from talking much more about it[d].

3. Protein-Protein interaction and 3D structure prediction

Graph based KNN is used in protein interaction prediction. Similarly KNN is used in structure prediction [d].

4 . KNN have been also called as an lazy folder because we have using it in an classifier field and it ranges an value we can look it in an classifier output[d]

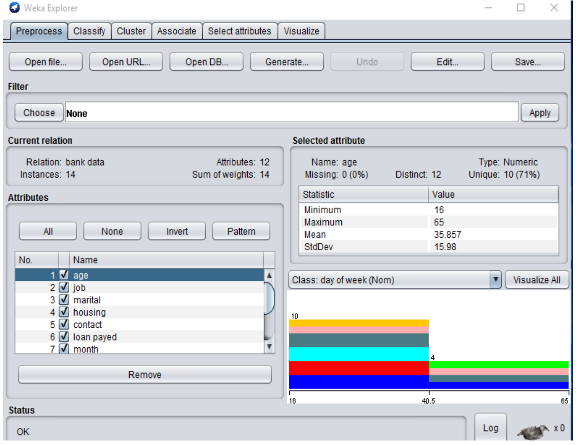
5. Bank credit risk assessment is widely used at banks around the world. since credit risk evaluation is very crucial ,variety of techniques is used for risk level assessment . In addition, credit risk is one of the main function of banking[5]

VI. Analysis Of KNN In Weka Using Banking Dataset:

Algorithm of using KNN in banking dataset using weka tool

Step1: Open the weka tool & in application field select explorer

Fig(A) explorer window



Filter: None

Current relation

Relation: bank data Attributes: 12
Instances: 14 Sum of weights: 14

Selected attribute

Name: age Type: Numeric
Missing: 0 (0%) Distinct: 12 Unique: 10 (71%)

Statistic	Value
Minimum	16
Maximum	65
Mean	35.857
StdDev	15.98

Attributes

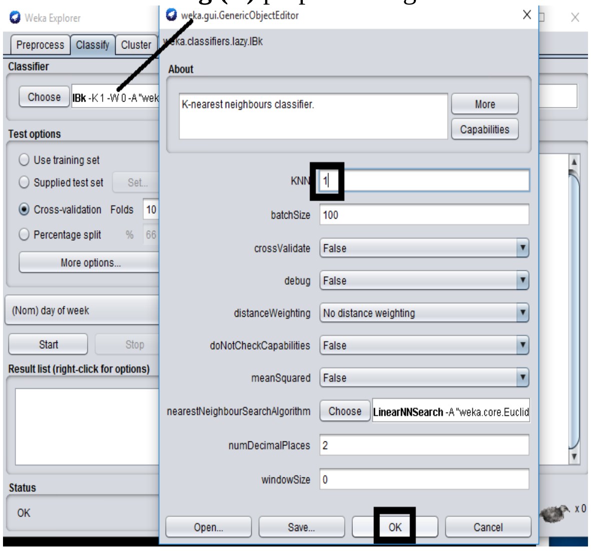
No.		Name
1	☑	age
2	☑	job
3	☑	marital
4	☑	housing
5	☑	contact
6	☑	loan payed
7	☑	month

Class: day of week (Nom)

Status: OK

Step2: In preprocess tab,select open file command in that open the banking dataset of extension csv, click edit button you can able to view the data. Click all attribute button

Fig (B) preprocessing



Classifier: IBk -K 1 -W 0 -A "wek

Test options: ○ Use training set ○ Supplied test set ● Cross-validation Folds 10 ○ Percentage split % 66

(Nom) day of week

Status: OK

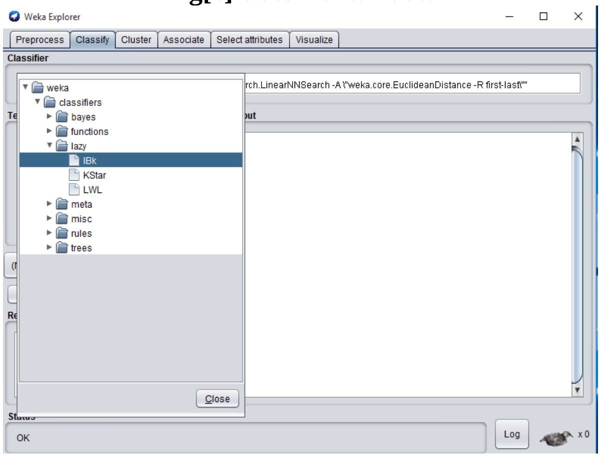
weka.classifiers.lazy.IBk

K-nearest neighbours classifier.

KNN: 1
batchSize: 100
crossValidate: False
debug: False
distanceWeighting: No distance weighting
doNotCheckCapabilities: False
meanSquared: False
nearestNeighbourSearchAlgorithm: LinearNNSearch -A "weka.core.Euclid
numDecimalPlaces: 2
windowSize: 0

Step3: Click classify tab, in that open the lazy folder in that ibk option

Fig[c] classifier window

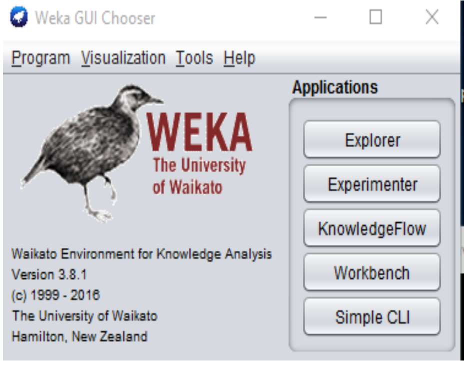


rch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

Status: OK

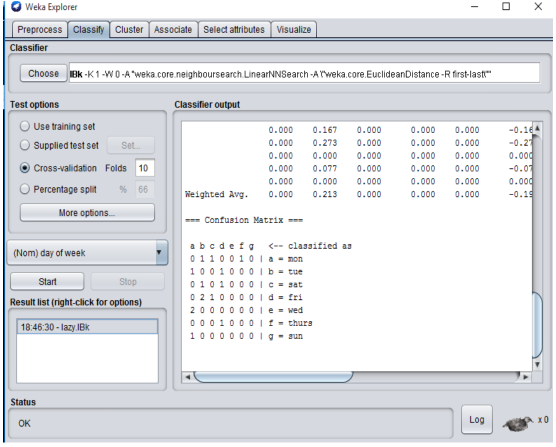
Step4: Click classifier field to look at the KNN neighborhood ranges as1 click ok(you can range the value as you like and note the values)

Fig (D) : changing KNN attributes in classifier field



Waikato Environment for Knowledge Analysis
Version 3.8.1
(c) 1999 - 2016
The University of Waikato
Hamilton, New Zealand

Step 5: Now you can note the tp rate ,fp rate, attributes,and note the confusion matrix.



IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) day of week

18:46:30 - lazy.IBk

```
                 0.000    0.167    0.000     0.000    0.000    -0.16
                 0.000    0.273    0.000     0.000    0.000    -0.27
                 0.000    0.000    0.000     0.000    0.000    0.000
                 0.000    0.077    0.000     0.000    0.000    -0.07
                 0.000    0.000    0.000     0.000    0.000    0.000
Weighted Avg.    0.000    0.213    0.000     0.000    0.000    -0.19

=== Confusion Matrix ===

 a b c d e f g   <-- classified as
 0 1 1 0 0 1 0 | a = mon
 1 0 0 1 0 0 0 | b = tue
 0 1 0 1 0 0 0 | c = sat
 0 2 1 0 0 0 0 | d = fri
 2 0 0 0 0 0 0 | e = wed
 0 0 0 1 0 0 0 | f = thurs
 1 0 0 0 0 0 0 | g = sun
```

OK

Fig (E) :confusion matrix

VII. Conclusion

In this paper it is discussed with basic ideas of Data mining ,Spatial,SpatioTemporal concepts. The nearest neighbor classifier can be regarded as a special case of the more general k-nearest neighbor's classifier.An analysis on KNN algorithm in density estimation ,Classification is list outed with its various applications.We have used banking datasets in weka tool for finding ranges valus. Thus Confussion matrix, tp,fp ratios are identified & classifications of Data is produced. As a final point a Synthetic data set is suggested to classify the customer details using size of the instance base, the performance of customers is the classified through K-NN, to extract the required statistical baseline, the assumption designed for all unknown instances belong to the class most frequently represented in the training data by means of KNN

References

[1] K. VenkateswaraRao, A .Govardhan, K.V .ChalapatiRao, "Spatial temporal Data Mining:Issues ,Task & Application",IJCSES",Vol 3.No,February 2012

[2] E.Baby Anitha,,Dr.K.Duraisamy, "Prediction of vehicle Movements using spatial Mining: A Recent survey",IJART,Vol.2 Issue 4,2012,pp-1-4.

[3] Bindiya, M.Varghese ,Unnikrishnan, A,Poulose, Jacob.K, Spatial clustering Algorithms-An Overview",AJCSIT,2013,1-8,ISSN 2249-5126

[4] gourav rahangdale, mr.manish agirwar, dr.mahesh motvani" IJCSI International journal of computer science issues",volume 13,issue 5,sep 2016

[5] aida krichene abdel moula "accounting and management information systems", volume 14,no.1,pp 79106,2015

[6] American Statistical Association, " journal of statistical software" November 2012,volume 51,issue 7 .