ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

Storm Identification in the Rainfall Data Using Singular Value Decomposition and K-Nearest Neighbour Classification

Manoj Praphakar.T [1], Shabariram C.P [2]

P.G. Student, Department of Computer Science Engineering, Sri Shakthi Institute of Engineering and Technology Coimbatore, India[1]

Assistant Professor, Department of Computer Science Engineering, Sri Shakthi Institute of Engineering and Technology Coimbatore, India[2]

ABSTRACT: In converting the rainfall data into valuable storm class is a challenging in terms to guarantee the quality of discovered relevance features in rainfall dataset for describing storm prediction large scale terms and data patterns. Most popular text mining and classification methods have adopted term-based approaches suffered from the problems of feature evolution and concept evolution. In this work, we present an innovative model for relevance feature discovery. It discovers rainfall conditions as higher level features and deploys them over low-level features (terms) using singular value decomposition. It also classifies and clusters the features of rainfall data into categories and updates term weights based on their specificity and their distributions in patterns using K Nearest Neighbour. Experimental results proves that proposed system outperforms in terms of f- measure precision and recall.

KEYWORDS: Storm analysis, TextMining, Rainfall, f-measure, K-NN, Precision, Recall

I INTRODUCTION

Big data is collection of large volumes of data that contains both the structured and unstructured, semistructed data which is difficult task to store analyze process, share, visualize and manage with the most modern traditional database and software techniques. Volume of data can be calculated by the total amount of transactions. Due to recent development increasing need on platforms and in many software industry applications to handle the scalability, accuracy, rate at which enterprises remain to face in a competitive global Market world .Major Big data challenges are capturing data, increase of storage capacities, transfer, searching, analysis, increase of processing power, presentation, availability of data. Along with traditional transactional and analytics data stores, are collected additional data across social media activity, web server log files, financial transactions and sensor data from equipment in the field. The sources of Big data fall in to three categories, they are

1. Streaming data
2. Social media data
3. Publicly available resources

II. RELATED WORK

Storm analysis model from Raw Rainfall dataset using techniques such as KNN and SVM has aims to predict the occurrence and strength of a storm by analyzing the rainfall data of that region. Most popular text mining and classification methods have adopted term-based approaches suffered from the problems of feature evolution and concept evolution. In this work, we present an innovative model for relevance feature discovery. It discovers rainfall

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

conditions as higher level features and deploys them over low-level features (terms) using singular value decomposition. It also classifies and clusters the features of rainfall data into categories and updates term weights based on their specificity and their distributions in patterns using K Nearest Neighbour.

The raw rainfall dataset is being trained by the SVM classifier .The trained dataset is then summarized into a model which performs the prediction of storm centric characteristics. Training process is implemented in Hadoop framework. We obtain a considerable improvement in the total performance of the system by employing KNN based classification. It is also used in the system for predicting the intensity of storm. In the existing system, the raw rainfall dataset is collected and stored in a relational database and then map-reduce based techniques are applied for storm analysis. In the proposed methodology as the raw rainfall dataset is being trained by KNN classifier the performance and accuracy rate got improved. Also, the training process is done on multi-node hadoop cluster by considering large raw rainfall dataset. With multi-node hadoop cluster there was a large reduction in the total training time. Storm depth of a particular region is calculated by applying Singular vector decomposition algorithm. This improved the total efficiency of the storm intensity prediction.

III. HADOOP

Hadoop is a open source software which is developed using Java programming that helps in accessing the large volumes of datasets in a distributed computing model. It is developed and managed by apache hadoop, Hadoop framework uses the MapRedue algorithm that helps the data is analysed in parallel. It stores any type of data in its own format and performs the analyses and changes on the data. Hadoop stores the information ranging from tera to even petabytes of data. It is efficient and reliable and handles the hardware failure automatically when the system malfunction occurs,with out any loss of data. The two components of hadoop are: 1)Hadoop Distributed File System(HDFS), 2) MapReduce. Both the HDFS and MapReduce are designed to continue to work in the face of System Failures. Hadoop runs code across a cluster of computers. Data are divided into directories and files.Files are divided into uniform blocks 128M and 64M. Servers can be added and removed from the cluster dynamically and hadoop operates to continue without any interruption.

IV. MAPREDUCE

Mapreduce is a functional programming model for data processing. Hadoop can run Mapreduce programs written in various languages namely Java, Ruby, Python, and C++. Mapreduce processing consists of two phases: the map phase and reduce phase. Each phase of Mapreduce consists of a key-value pairs as input and output. Hadoop divides the input to a Mapreduce job into the fixed-size pieces called input splits. The Mapreduce operations is based on shuffle, sort and reduce.

Functions	Input	Output
Map	< k1, v1 >	list (< k2, v2 >)
Reduce	< k2, list(v2) >	list (< k3, v3 >)

Table 3.1 Input Output types of MapReduce job

Map() Function: It performs filtering and sorting of task into queue.
Reduce() Function: It performs a summary operation of best candidate resource for task execution.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

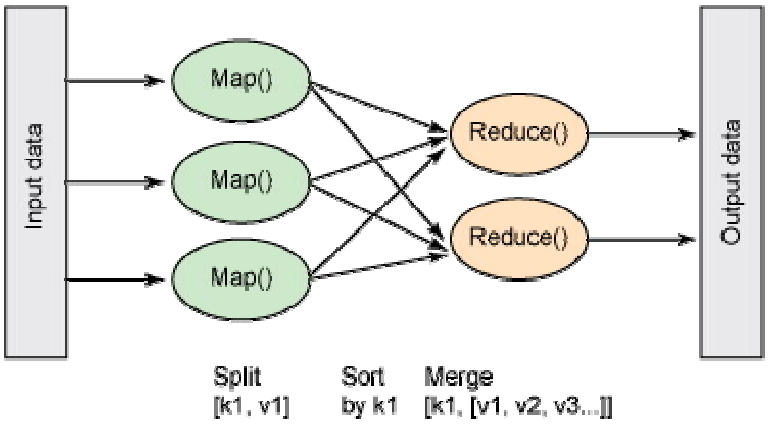


Figure: 3.2 Dataflow in Mapreduce

Analysing the National Weather Service (NWS) Dataset

The NWS is an agency of United States that provides the weather forecasts and other storm related warnings given to organizations for their prior purpose of protection against the disasters. They provide the detail information of data(ie) from a period of twenty years. Following are the dataset parameters that are required for predicting the storm related data. The is available from National Climate Data Center (NCDC, http://www.ncdc.noaa.gov/)

V. IMPLEMENTATION

1. Model Feature Reduction technique using Singular value decomposition

The Feature reduction is to decrease the processing time required to perform a classification and improve overall classification accuracy. Transforming the input data into the set of features, If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. Feature extraction is performed on raw data prior to applying *k*-NN algorithm on the transformed data in feature space. To extract only the meaningful concepts from the big raw rainfall data by filtering out unnecessary/excess data without trading off the small limited area of analysis.

2. Classification of the Rainfall data using K nearest neighbour for Storm identification

The storm Identification is carried out using K-NN algorithm, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. The process consists of four main components: event separator, location proximity creator, sub storm identification, and main storm identification. Event separator takes relational data as an input and identifies local storms for a particular site. Event separator uses selection and sorting database features to complete the process. The sub storm identification takes location proximity and local storms as inputs and identifies hourly storms at each hour. The curse of dimensionality in the *k*-NN context basically means that Euclidean distance is unhelpful in high dimensions because all vectors are almost equidistant to the search query vector (imagine multiple points lying more or less on a circle with the query point at the center; the distance from the query to all data points in the search space is almost the same). Finally, main storm identification combines consecutive hourly storms that meet grouping- and spatial- windows requirements to create the overall storms.

3. Normalization of the Dataset

This is the phase in which both the normalization and training of dataset is performed. Data mining approaches we need to normalize the inputs; otherwise the network will be ill conditioned. In essence, normalization is done to have

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

the same range of values for each input to the KNN Model This can guarantee stable convergence of weights. In the training process the correct output for each input record is known and the output nodes are assigned with these correct values. These error terms are then used to adjust the weights in the hidden layers so that during next iteration the output values will be closer to the "correct" values

VI. SYSTEM ARCHITECTURE FOR PROPOSED WORK

(a) Working of Map function

- In Classification Process, Mapper function is designed as classes and data are sorted based on constraints – k values
- Mapper function is defined as job with location and size of input data
- Assign weights to the neighbors based on their 'distance' from the query point
  - Weight may be inverse square of the distances
- Hadoop provides that engine through (the file system we discussed earlier) and the JobTracker + TaskTracker system.
- TaskTracker monitors the execution of the task (Data attribute of rainfall dataset)and updates the JobTracker through boundary or Target value .
- Intermediate merging on the nodes are also taken care of by the JobTracker
- After Tracking, it display the feature vector of the reduced Attributes in the rainfall dataset

(b). Working of Reduce Function

- It aggregates the Map jobs and Classify it
- JobTracker is simply a scheduler.
- TaskTracker is assigned a Map or Reduce (or other operations);
- Map or Reduce run on node and so is the TaskTracker;
- Each task is run on its own JVM on a node
- It classifies the data attributes based on Class generated through target functions

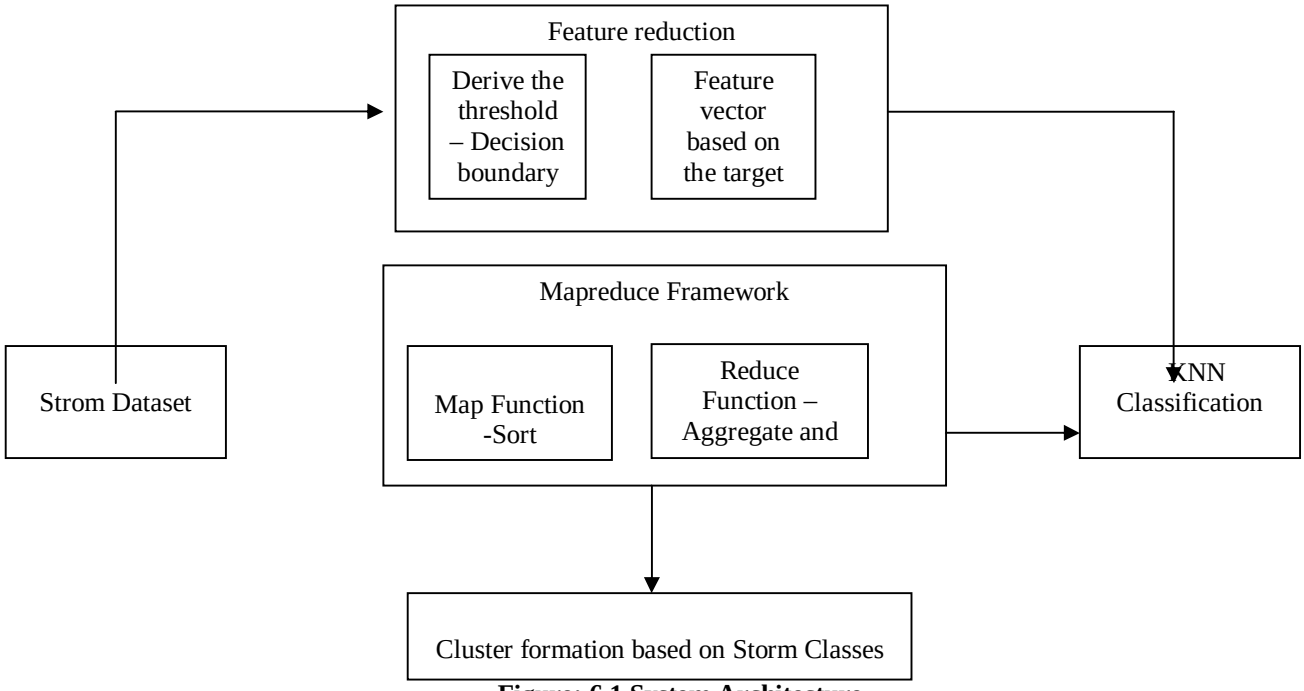


Figure: 6.1 System Architecture

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

VII. PERFORMANCE EVALUATION

The performance of the KNN based Storm Classification is computed using

(a) Euclidean Distance – It is distance between the two data points in the cluster.
(b) Precision - is the fraction of retrieved instances that are relevant in the cluster
(c) Recall - is the fraction of relevant instances that are retrieved.

Storm classification Based on the K –NN constraints

- KNN is a classifies the dataset based on the available attributes and based on a similarity measure
- It iterate periodically , Hence the Iterative procedure is as follows
- If K=1, select the nearest neighbor , where k can be Attribute or value of rainfall dataset
- If K>1, For classification select the most frequent neighbor.
- For regression calculate the average of K neighbors.
- Features extracted by the K-NN classifier are all as follows
- All instances correspond to points in an n-dimensional Euclidean space
- Classification is delayed till a new instance arrives
- Classification done by comparing feature vectors of the different points
- Target function may be discrete or real-valued.

Classification technique	Precision	Recall	F measure	Cluster size
SUPPORT VECTOR MACHINE	0.7	0.3	0.42	4
KNN	0.6	0.4	0.52	2

Table 2: Performance Evaluation of the Proposed System

The Above table explains the performance of the proposed (KNN) and Existing (SVM) against the precision, Recall and f- Measure.

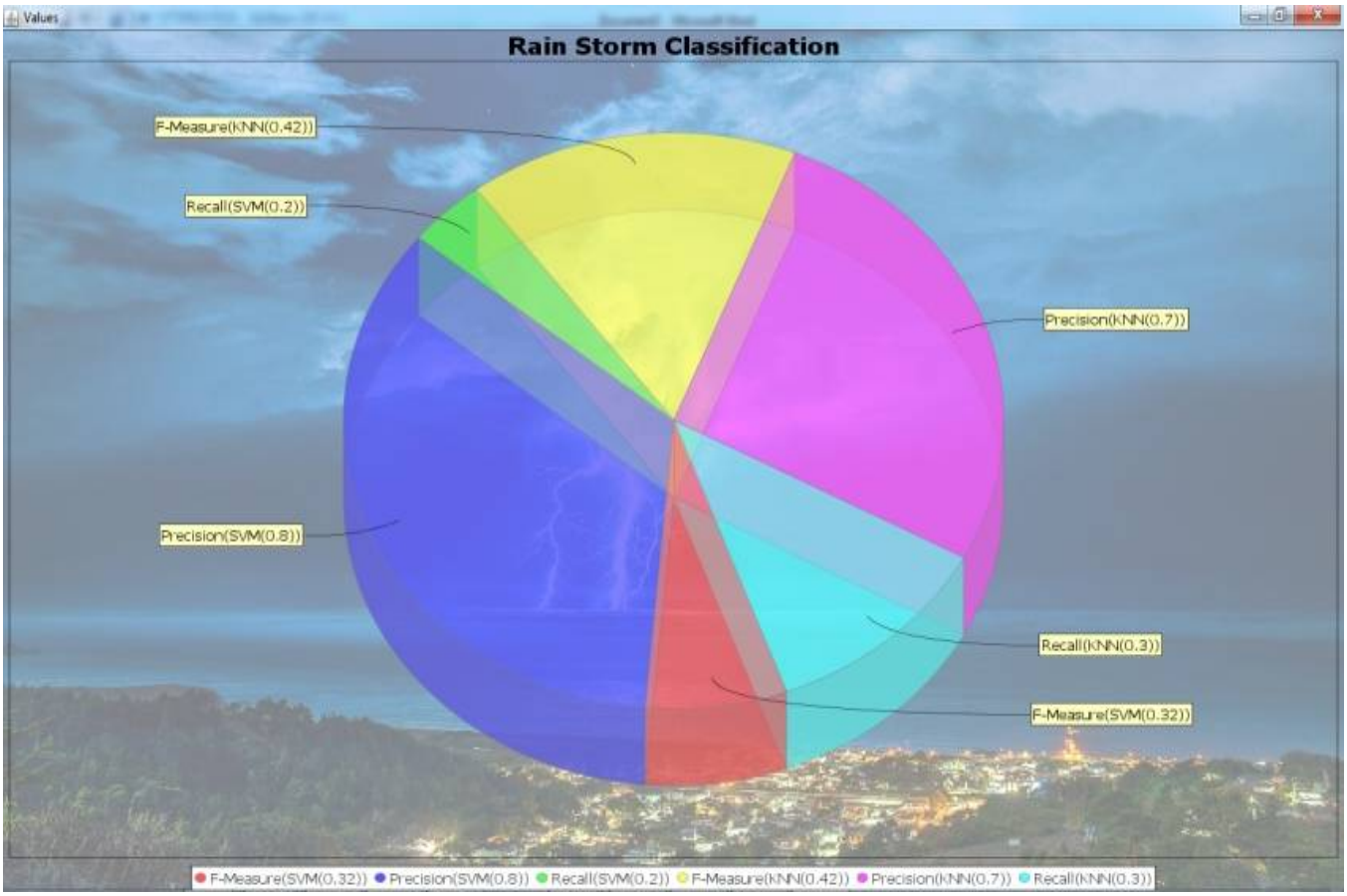
The following figure also depicts the performance of the proposed system with respect to several metrics.It also classifies and clusters the features of rainfall data into categories and updates term weights based on their specificity and their distributions in patterns.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016



VIII. CONCLUSION

In this work, designed and implemented a model for relevance feature discovery. It discovers rainfall conditions as higher level features and deploys them over low-level features (terms) using singular value decomposition. It also classifies and clusters the features of rainfall data into categories and updates term weights based on their specificity and their distributions in patterns using K Nearest Neighbour. The Data classification is carried in the map reduce paradigm using Hadoop framework as the dataset is available in large scale and hence in order to improve the performance of the cluster scalability, it has been utilized to classify the rainfall data into cluster using the Mapper and reduce functions.

REFERENCES

[1]. K. Jitkajornwanich, R. Elmasri, C. Li, and J. McEnery, "Extracting Storm-Centric Characteristics from Raw Rainfall Data for Storm Analysis and Mining," Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (ACM SIGSPATIAL BIGSPATIAL'12), 2012, pp. 91-99.
[2]. K. Jitkajornwanich, U. Gupta, R. Elmasri, L. Fegaras, and J.McEnery, "Using MapReduce to Speed Up Storm Identification from Big Raw Rainfall Data," Proceedings of the 4th International Conference on Cloud Computing, GRIDs, and Virtualization (CLOUD COMPUTING'13), 2013, pp. 49-55.
[3]. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proceedings of the 6th Symposium on Operating Systems Design and Implementation (OSDI'04), 2004.
[4]. C. Lam, Hadoop in Action. Dreamtech Press, New Delhi, 2011.
[5]. M. Young, The Technical Writer's Handbook. Mill Valley, CA: UniversityScience,1989.
[6]. R. Elmasri and S. Navathe, Fundamentals of Database Systems, 6th ed. Pearson Education, Massachusetts, 2010.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 5, May 2016

[7]. A. Overeem, T. A. Buishand, and I. Holleman, "Rainfall Depth-Duration-Frequency Curves and Their Uncertainties," Journal of Hydrology, vol. 348, 2008, pp. 124-134.
[8]. W. H. Asquith, M. C. Roussel, T. G. Cleveland, X. Fang, and D. B.Thompson, "Statistical Characteristics of Storm Interevent Time,Depth, and Duration for Eastern New Mexico, Oklahoma, and Texas," Professional Paper 1725. U.S. Geological Survey, 2006.
[9]. W. H. Asquith, "Depth-Duration Frequency of Precipitation for Texas," Water-Resources Investigations Report 98-4044. U.S.Geological Survey (USGS), 1998.
[10] Virginia Department of Conservation and Recreation, "Stormwater Management: Hydrologic Methods," retrieved: May 2, 2012, from: ttp://dcr.cache.vi.virginia.gov/stormwater_management/documents/Chapter_4.pdf.