Evaluating renewable energy using data mining techniques in developing India

Anushree A. Wasu,Harshada M .Kariya,Shreyas S. Tote

Abstract- Developing country like India expects foreign investments in industrialization to achieve economic growth. This paper deals with the study of opportunities, challenges in using renewable energy resources in India and applying data mining algorithms in the renewable energy source data. The daily temperature values of a city or place acts as data for prediction of rainfall, energy calculation etc. There are weather stations available which records the weather of almost all the region in the earth. The weather data acquired in these stations are raw data and found in abundance. The data mining techniques can be applied to these raw data to acquire meaningful patterns out of it, which can be used to predict rainfall, solar energy availability etc.. These algorithms cover classification, clustering, statistical learning, association analysis, and link mining, which are all among the most important topics in data mining research and development.

Index Terms— Renewable energy sources, Non-renewable energy sources, Weather Data, Data Mining, Clustering, Simple K-Means, Expectation Maximization, Solar energy.

1 INTRODUCTION

The Indian Subcontinent belongs to the tropical region and has approximately 250 to 300 sunny days out of 365 days a year .Therefore there are high opportunities of acquiring solar radiation from Indian regions for the use of production of solar energy. Using the temperature values the sun shine of these cities are calculated. The clustering algorithms namely simple k-Means algorithm and expectation maximization algorithm are used to compare the sun shine of these cities which in turn facilitates in know in the amount of solar radiation that can be acquired in the above mentioned cities. The data mining tool used in this paper is Weka, an open source data mining tool. is strong support for promoting renewable sources such as solar power and wind power, requiring utilities to use more renewable energy (even if this increases the cost), and providing tax incentives to encourage the development and use of such technologies.

Data mining is the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources. Data mining techniques are the result of long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery.

2. ENERGY SOURCES

2.1 Non-Renewable Energy Sources

The Non-renewable energy sources are sources which are available on the earth and that cannot be re-generated within a short span of time. The Non-renewable energy sources include coal, oil, natural gas, nuclear power etc. The Non-renewable energy has some advantages that make them viable in country like India. They are cheap and easy to use. A small amount of nuclear power can be used to produce large amount of power. The Nonrenewable energy sources also have serious effects on environment that they are prone to exhaust and pollute the environment during the production of energy out of them. In India, the mainly used energy source is coal. Due to deficiency of domestic supply of coal to meet the needs of energy consumption, India imports coal from foreign countries which has raised the production cost of energy. These factors about Non-

- *Anushree A. Wasu is currently pursuing bachelor degree program in computer science and engineering in Amravati University, India, PH 8378803550. E-mail: anushreewasu@gmail.com*
- *Harshada M. Kariya is currently pursuing bachelor degree program in computer science and engineering in Amravati University, India, PH-9420022259. E-mail: Harshadakariya@yahoo.com*
- *Shreyas S. Tote is currently pursuing bachelor degree program in computer science and engineering in Amravati University, India, PH-8087961065. E-mail: shreyastote@yahoo.com*

renewable energy sources make us to move towards the alternative (i.e.) the renewable energy sources. The renewable energy sources have better opportunities for power consumption and do not have serious effects on the environment.

2.2.Renewable Energy Sources

India is a choice of many developed countries to expand their industrial sector which provides employment opportunities for people. The Indian Government focuses on providing the sophisticated environment with all possible facilities. The Renewable energy is energy which comes from natural resources such as sunlight, wind, rain, tides and geothermal heat which are renewable(naturally replenished).The Indian Government has formed a "Ministry of New and Renewable Energy", as an initiative towards the production of energy from renewable energy sources. The broad aim of the Ministry is to develop new and renewable energy sources to supplement the energy requirement of India . The Ministry also supports various projects in renewable energy sectors.

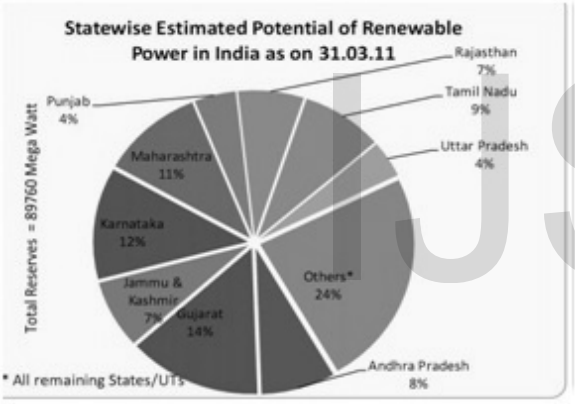

Fig 1. State wise estimated potential of Renewable power in India .

2.2.1 Wind Power

The most important form of renewable energy India depends on as alternate is Wind Power. The development of wind power industry in India took place in 1990s and India is the fifth largest installed wind power in the world . The states in India which contribute to this production are Tamil Nadu, Gujarat, Maharashtra, Karnataka, Rajasthan, Madhya Pradesh, Andhra Pradesh, Kerala, Orissa and West Bengal. The Wind Resource Assessment Program which is being coordinated by the Centre for Wind Energy Technology(C-WET) has so far covered 31 states and Union Territories involving establishment of about 1244 wind monitoring and wind mapping stations.

2.2.2 Solar Energy

Solar energy is one of the important renewable energy sources available in India. The solar energy is obtained from the sun and can be easily absorbed and converted into electrical energy for usage. There is some awareness among the people about the energy sector, increase in price of the energy sources and the alternate ways to be adopted. The people have also started to conserve solar energy and use it. Solar energy has a wide variety of applications in India.

2.2.3 Biomass Power

Biomass is plant matter used to generate electricity with firewood, animal dung; bio-degradable, waste from cities and crop residues is a source of energy when it is burnt. Biomass does not add carbon dioxide to the atmosphere which is considered to be a pollutant. Biomass is used as an alternative option in many rural areas in India.

2.2.4 Climate data

The climate is defined as the statistical information that describes the variation in weather at a region for as specified interval of time. Where as weather data is different that it describes the temperature, humidity, wind for a short interval of time at specified interval. The type of data used in this paper is climate data that is of the year1901 to 2000. Normally the temperature variations are affected only over a time period of 100 years or more, so the data set is sufficient to make the study.

2.2.5 Solar Radiation

Solar radiation is the radiant energy emitted by the sun. The solar emitted by the sun is acquired by the solar panels to produce solar energy out of it. Solar energy is a form of renewable energy. The sun is inexhaustible form of energy which can be used by the tropical regions. The solar energy can be stored and used as an alternative fuel. The amount of solar radiation reaching the Earth depends on various factors such as time, location, season etc. When the Earth is nearer to the sun the amount of solar radiation received by the Earth is high. Therefore it is necessary to check all the parameters such as location; weather etc when there is a plan of implementation of solar energy production system. The intensity of the solar radiation reaching the earth is approximately 1369 watts per square meter [W/m2]. This value is called as the Solar Constant. The total solar radiation (R) is calculated with the
formula

$$R = S\pi r2 / 4\pi r2$$

Where S is the Solar Constant in W/m2 and r is the Earth's radius. The surface weather stations which record the daily radiation are less compared to the weather stations which records temperature, humidity etc. Therefore the temperature values can be used to estimate the daily solar radiation. In this seminar we have taken the monthly maximum mean temperature values to calculate the months in which the maximum sunshine is recorded in the four cities namely Chennai, Coimbatore, Madurai and Kanyakumari of South India.

3. IMPORTANCE OF DATA MINING TECHNIQUES

Before implementing a technology among people, there have to be social acceptance to implement it. Though many parts of rural and urban India have the awareness of global warming, there is only little/ no awareness about the shortage in non-renewable energy sources and price increase in production of it. Proper planning and integration is another aspect to be considered. The renewable energy is source dependent such as water, air, sun etc so the plant has to be planned in places of its availability.

Rapid computerization of businesses produce huge amount of data. How to make best use of data? A growing realization: knowledge discovered from data can be used for competitive advantage.

4.THE DATASET

4.1 Data collection

Weather data is of two types one is synoptic data which are real time data which are provided for aviation safety and forecast modeling. The other type of data is climate data in which some kind of quality control is performed on it. The data set consists of monthly mean maximum, minimum temperature and total rainfall based upon 1901- 2000 data of various cities in India. We have taken the data of four cities namely Chennai, Coimbatore, Madurai and Kanyakumari. The cities were selected upon the values of latitudes and longitudes which play an important role with their climatic conditions. The latitude and the longitude of the city are given in the table 1.

City	Latitude(in degrees)	Longitude(in degrees)
Chennai	13.0810 N	80.2740 E
Coimbatore	11. 0183 N	76.9725 E
Madurai	9.9300 N	78.1200 E
Kanyakumari	8.0800 N	77.5700 E

Table I. Latitude and Longitude value

4.2 Description of the Dataset and pre-processing

The attributes present in the data set are Station name, Month, Period (years), Number of years, Maximum Mean Temperature in degree Celsius, Minimum Mean Temperature in degree Celsius and Mean Rainfall in Millimeter. Data pre-processing is the step in data mining, which cleans the data to improve its quality to meet the needs of the application. Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. Considering the requirement of the data for this seminar, the pre-processing of data was done by selecting only the needed attributes for clustering.

5 DATA MINING TECHNIQUES

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases.

5.1 Clustering

The Clustering method is define as the grouping of similar type of data. This helps in learning the number of similar kind of data available in a dataset. The clustering takes a seed value and depending on it, it clusters the data into clusters. The clustering algorithms work on the principle of measuring the similarities between the given set of objects by finding the distance between each pair. There are different kinds of distance methods being available such as Euclidean distance, Manhattan distance, Chebychev and categorical data distance. All the distance does not suit for all the clustering algorithms. So the choice of the distance depends on the clustering algorithm used.

The different types of clustering methods available are Partition Method, Hierarchical Method, Density-based Method, Grid-based Method and Model-based Method. The algorithms we have used in this paper are Simple k-Means and Expectation Maximization algorithm which comes under the partition method of clustering. The partition method is based on the greedy heuristics in which they are used in iterative manner to obtain a local optimum solution. A few good reasons ... 1)Simplifications 2)Pattern detection 3)Useful in data concept construction 4)Unsupervised learning process

The Clustering method is described as:

1. Begin with the disjoint clustering having level L(0) = 0 and sequence number m = 0.
2. Find the least dissimilar pair of clusters in the current clustering, say pair (s) according to d[(r),(s)]=mind[(i),(j)]
   where the minimum is over all pairs of clusters in the current clustering.
3. Increment the sequence number : m = m +1. Merge clusters (r) and (s) into a single cluster to form the next clustering m. Set the level of this clustering to L(m) = d[(r),(s)]
4. Update the proximity matrix, D, by deleting the rows and columns corresponding to clusters (r) and (s) and adding a row and column corresponding to the newly formed cluster. The proximity between the new cluster, denoted (r,s) and old cluster (k) is defined in this way: d[(k), (r,s)] = min d[(k),(r)], d[(k),(s)]
5. If all objects are in one cluster, stop. Else, go to step 2.

5.2 Simple k-means algorithm

The simple k-means algorithm is one among the widely used clustering algorithm. It is also called as centroid method as in each step, the centroid value of each cluster is assumed to be known and each point is allocated to the cluster depending on the distance between the centroid of the cluster and the point The algorithm used Euclidean distance method for distance calculation The simple k-means method is described as:

1. Select the number of clusters (k).
2. Assume k seeds as centroids of the k clusters. The seeds can be randomly chosen by the user if the values of data are unknown.
3. Compute the Euclidean distance of each object of the dataset from each of the centroids.
4. Allocate each object to the cluster if the distance between the centroid of the cluster and the object is small.
5. Compute the centroids of the clusters by calculating the means of attribute values of the objects in the cluster.
6. Stop the algorithm if the stopping criterion is met or go to step 3.

In this paper the simple k-means algorithm is computed using the open source software tool Weka.

5.3 Expectation maximization (em) algorithm

The expectation maximization algorithm works in contrast to the simple k-means algorithm. The expectation maximization algorithm works on the concept of assuming that the objects in the dataset have attributes whose values are distributed. The Expectation Maximization described as:

1. Assume the initial values
2..Use the normal distributions and calculate the probability of each object belonging to the two clusters.
3. Calculate the possibility of data coming from the two clusters.
4. Iterate the process by re-assuming the parameters and go to step 2 until the stopping criterion is met. The Expectation Maximization algorithm assumes that the attributes involved are independent and normally distributed.

6 ILLUSTRATION WITH EXAMPLE

The dataset was loaded in the Weka environment and the data was pre-processed by attribute reduction. The simple k-means and Expectation Maximization algorithm were applied. The Weka environment is shown in Figure2.The Weka environment has four applications namely Explorer, Experimenter, Knowledge Flow and Simple CLI. The dataset collected was converted into file with extension of .arff and loaded into the Weka environment .The simple k-Means algorithm formed two clusters in which the maximum temperature was recorded in the month of May in the four cities namely Chennai, Coimbatore, Madurai and Kanyakumari. The maximum temperature in turn implies the maximum of sunshine hours or daylight in which the solar radiation acquired will be high. The Chennai city recorded the maximum of monthly

Fig 2. The Weka Environment

mean temperatures compared to the other cities. The Expectation Maximization algorithm formed four clusters, in which the maximum of monthly mean temperature was recorded in the month of May and June.

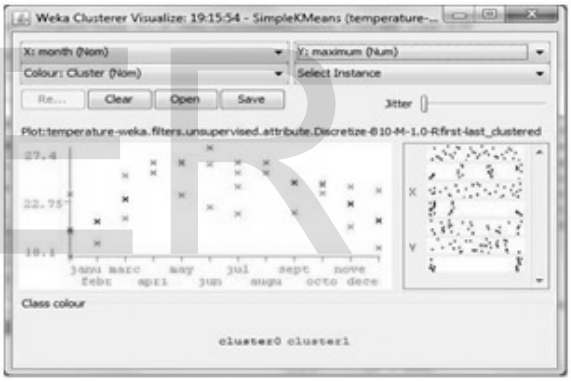

Fig .3 The simple k-means clusters visualization

The visualization of the clusters formed by the two algorithms is given in the figures 2 and 3.

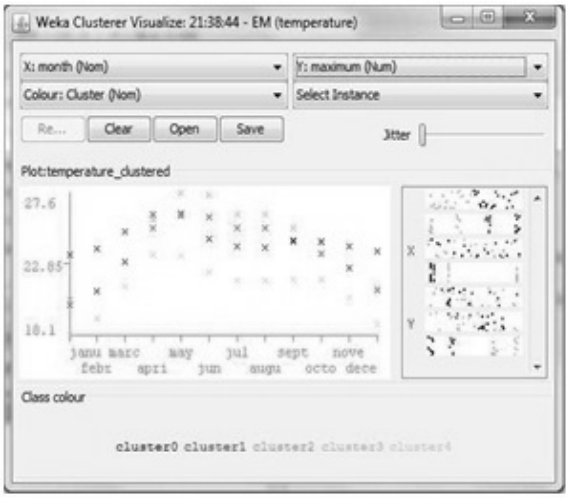

Fig.4 The EM clusters visualization

7 CONCLUSION AND FUTURE WORK

The dataset of the monthly mean maximum and minimum temperature of four cities Chennai, Coimbatore, Madurai and Kanyakumari were collected and clustering algorithms simple k-means and expectation maximization algorithm were applied. The clusters formed indicated that the maximum monthly mean temperature was recorded in the month of May and June which implied the maximum sunshine hours in these months. The maximum in temperature implies the maximum daylight or sunshine hours. The sunshine hours determines the amount of solar Radiation that can be acquired. The maximum amount of sunshine is recorded in the city of Chennai when compared In future They can also be used to check the feasibility of the installation of the plant and the prediction of the energy to be produced in it. The available data mining algorithms can be adapted to our systems with needed modifications.

ACKNOWLEDGEMENT

We sincerely convey our thanks to our guide Professor P.A Patil for her guidance, support and help in each and every aspect .The authors are grateful for the constructive comments of the earlier theory of Data Mining Techniques

REFERENCES

[1] Estimating the availability of sunshine using data mining techniques (ICCCI 2013) Jan 4-6 2013
[2] M.Mayilvahanan,M.Sabitha,Opporunities and challenges in using renewable resources in India: A data mining approach, International Journal of Emerging Trends and Technology in Computer Science, Volume 1, Issue 2, July-August 2012.
[3] J.Han and M.Kamber, Data mining: concepts and techniques, Morgan Kaufmann, 2012
[4] Mark Hall, Eib Frank, Geoffrey Holmes, Bernhard Pfahringer, PeterReutemann, Ian H.Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.
[5] Jin R, Goswami A, Agrawal G (2006) Fast and exact out-of-core and distributed *k*-means clustering. KnowlInfSyst 10(1):17–40
[6] Dhillon IS, Guan Y, Kulis B (2004) Kernel k-means: spectral clustering and normalized cuts. KDD 2004, pp 551–556
[7] Peter E. Thornton, Steven W.Running, An improved algorithm for estimating incident daily solar radition from measurements of temperature, humidity, and precipitation, Agricultural and Forest Meteorology, Elsevier, pp. 211-228,1999.
[8] Wenke Lee and Sal Stolfo. ``Data Mining Approaches for Intrusion Detection'' *In Proceedings of the Seventh USENIX Security Symposium (SECURITY '98)*, San Antonio, TX, January 1998
[9] Moon, T.K, The expectation-maximization algorithm, Signal Processing Magazine, IEEE Magazine,1996.
[10] Jain AK, Dubes RC (1988) Algorithms for clustering data. Prentice-Hall, Englewood Cliffs
[11] Nations Environment Programme Global Trends in Sustainable Energy Investment 2007: Analysis of Trends and Issues in the Financing of Renewable Energy and Energy Efficiency in OECD and Developing Countries (PDF), p. 3.