International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 9717-9722

Classification of Rainfall Data Using Linear Kernel Based Support Vector Machines

Kolluru Venkatanagendra

Research Scholar, Department of Computer Science,
Vikrama Simhapuri University, Kakutur, Nellore,
SPSR Nellore, Andhra Pradesh, India.

Orcid Id: 0000-0001-6013-3209

Dr. Maligelaussenaiah

Assistant Professor, Department of Computer Science,
Vikrama Simhapuri University, Kakutur, Nellore,
SPSR Nellore, Andhra Pradesh, India.

Abstract

Classification is a data mining technique used to predict group membership for data instances.Exploration of satellite imagery has become increasingly important in several application domains such as change detections, fire risk mapping etc... There are many algorithms used for image classification these days. In this paper we will be examiningfew popular algorithms used for classifying remote sensed data like decision tree classifier, Multi-layer perceptron classifier, Support Vector machines and Naïve Bayes classifier. To optimize the classification accuracy of support vector machines an algorithm will be proposed. After that we will discuss the performance of these algorithms depending on different parameters and comparing their correct rate in different categories.

Keywords: Classification, Data mining, classifier, Support Vector Machines, SVM.

INTRODUCTION

Eliciting vital information from huge amounts of repositories of data in any organization is a very important component of building organization strategy. The process of accomplishing the same is known as data mining. It encapsulates the process of amalgamating variety of data from diverse sources and eliciting information from the same [1].

Information can be elicited from data in two main ways. The first one is called supervised learning. This approach is used when data can be categorized into two or more categories. Data is first separated into training and test sets. Using training set the model is trained and the same is validated against the test set. Once the model is thus built the same can be used to predict the category of any new data set. Another approach is used when the data cannot be categorized into two or more categories. It is called unsupervised learning. In this paper the approach is restricted to discussing supervised learning [2].

In this paper we shall be investigating Naïve bayes, decision tree, MLP and SVM algorithms. The remote sensed data that we shall be using in this collected from the IMD portal. The rainfall dataset classification we shall discuss in this paper is useful in fisheries oceanography, biological, physical, marine geology and coastal ecology.[3-4].

In this research paper the classification of rainfall data is done using SVM for which diverse characteristics are exploited to portray the evidence level confined in the data set. The suggested technique is then matched with Naïve bayes, decision tree, multi-layer perceptron classifiers. Various statistical parameters are used to compare these methods. The result of comparing these methods is that we lead to the conclusion that the SVM method is the best compared to the other used methods on the said data set[5].

Our next section presents the proposed methodology. Section III describes Background Knowledge about various classifiers discussed and in section IV discussion of results is carried out. Section V summarizes the work and in section VI data source is acknowledged and in the final section references are given.

PROPOSED METHODOLOGY

The approach used in this paper has many steps as illustrated in the below mentioned diagram:

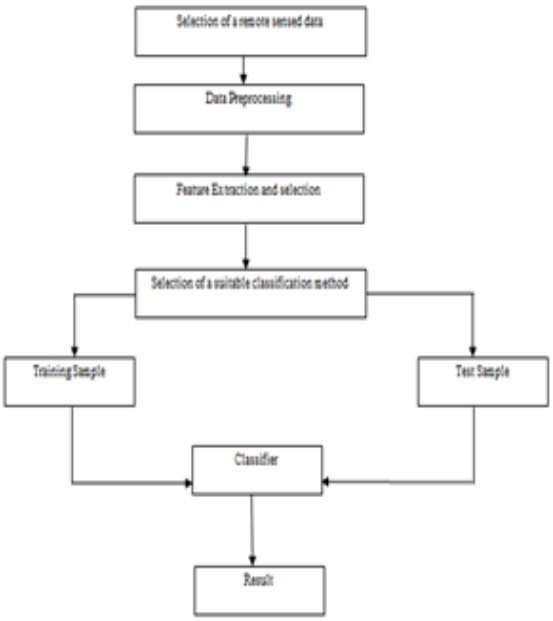


FIGURE 1: PROPOSED MODEL

Data Selection

The first obvious step in any data mining procedure is data selection. Hence in this paper data is selected from the IMD portal.

Processing of data

The data thus collected shall have many erroneous entries like missing values, duplicate values etc. The same is first cleaned for removing data anomalies.

Feature Extraction and Selection

Feature selection happens to be an important step in a supervised learning process that shall be used in this paper.

Supervised classification approach selection

Different supervised classification approaches have diverse merits and demerits[8]. The selection of an appropriate classification approach is an art rather than a science[17-18].

In this paper we shall be using SVM classification method to classify the rainfall datasets and shall be comparing its performance with Multi Layer Perceptron, Naïve Bayes, and Decision Tree classification methods.

Background Knowledge the mentioned classification methods are given in Section III.

Training and Testing

The obtained model from the previous method is validated against the test set[19].

BACKGROUND KNOWLEDGE

Naive Bayes

This classification method is grounded on the famous concept called Bayes rule. It has an underlying assumption that attributes x and y are conditionally independent. [6-7].

In this approach a data item x is classified into a class with maximum posterior probability and also if it satisfies the below equations[8].

$$P(P_i/Q)>P(P_j/Q) \text{ for } 1<=j<=m, j\neq i \quad (7)$$

$$P(P_i/Q) = \frac{P(Q/Pi)P(Pi)}{P(Q)} \quad (8)$$

If the probabilities cited herein are unknown, then,

$$P(P_1)=P(P_2)=\ldots=P(P_n) \Rightarrow \text{maximize } P(Q|P_i) \quad (9)$$

Class prior likelihoods can be assessed by $P(P_i)=|P_i,M|/|M|$

Given $Q(q_1,\ldots,q_n)$, $P(Q|P_i)$ is:

$$P(Q/P_i) = \prod_{k=1}^{n} P(\frac{qk}{Pi})$$

$$= P(q1/P_i)qP(q_2/P_i).....xP(q_n/P_i) \quad (10)$$

The probabilities $P(q_1|P_i)$, $\ldots P(q_n|P_i)$ can be estimated from the training tuples[9-11].

MLP Classifier Over View

This approach makes use of a layer of neurons wherein output from one layer is supplied as input to another layer of neurons and Euclidean distance is used to minimize the cost as in Equation(11)

$$E(w) = \frac{1}{2}\sum_{i=1}^{N} ||y(ai,b) - di||^2 \quad (11)$$

Where a is input and b is output vector [7].

As shown in Fig. 1. The MLP layer mentioned here is built with 3 layers wherein every neuron has its own weight and bias[9].

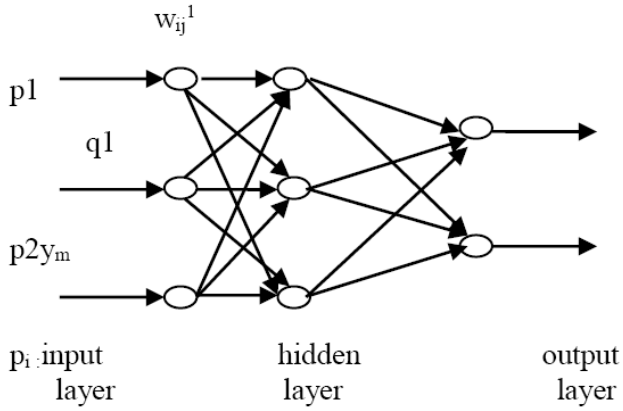


Figure 1: Simple Neural Network.

Any layers input is attained from weighted entirety of preceding layer

$I^J_j=\sum_{k=1}^{Nk} \quad WjkO_k^{J-1}$ (12)

and

$p_k^{J-1}=\phi(Q_k^{J-1})$ (13)

where O is output and P is input and j and k are layer numbers[15-17].

Overview of Decision tree Classifier

Decision trees are classically built recursively, following a top-down approach. The facts attained in the learning process is symbolized as a tree wherein each internal node comprises a query with respect to one precise attribute (with one offspring per potential answer) and each and every leaf is labeled with one of the probable classes. While examining decision tree one initiates at the root node and then follows the replies to the various possible questions in the internal nodes until and unless a leaf node is reached[18]. Classification tree is built in consensus with splitting rule that undertakes the splitting of learning sample into minor parts. A measure for tree splitting t is grounded on a node impurity function I(t)[20-22].In this paper Conditional Inference trees were used to classify the said datasets.

Overview of the SVM

SVM is a supervised technique wherein data is mapped onto a higher order space using a kernel function: $K(x,xj)=\langle\phi(x),\phi(xj)\rangle$. Decision function can be statedas[23-24] (1):

$$f(x) = \sum_{j=1}^{Sv} \alpha jyjK(x, xj) + b \qquad (1)$$

αjisLagrange multipliers, K is kernel function and b is the bias which is calculated using a support vector[25]. Then, the optimum hyper-plane relates to f (x)=0. Henceforth, test data can be dented as (2):

$$x\epsilon\begin{cases} positivecategoryiff(x) > 0 \\ negativecategoryiff(x) < 0 \end{cases} \qquad (2)$$

The specified training set of case label pairs as delivered by equation[26] (3).

$$(x, y) = \{ (x_1,y_1), (x_2,y_2), \ldots\ldots (x_n,y_n)\} \qquad (3)$$

Where $x_n\epsilon R^y$ and $y_n\epsilon$ {-1, +1} and SVM desiresresolution of the optimization problem$\frac{1}{2}W^TW + \sum_{i=1}^{l}\xi_i$C

Subject to as given in equation (4)

$$y_i\ (w^T\Phi(x_i) + b) \geq 1\text{-}\xi_i\xi_i \geq 0 \qquad (4)$$

In this paper 740 support vectors classify rainfall data into two sets. The training error in the said classification process was 0.123109.

RESULTS AND DISCUSSIONS

The algorithms thus discussed is applied for classifying the rainfall datasets into two groups that is districts with more rainfall and districts with less rainfall while in the end performance of classification is measured by means of diverse metrics. R software is used to perform the mentioned operations [27].

Data Acquisition

Around154 records of rainfall data from various districts from various districts of Andhra Pradesh are obtained.

Features Extraction

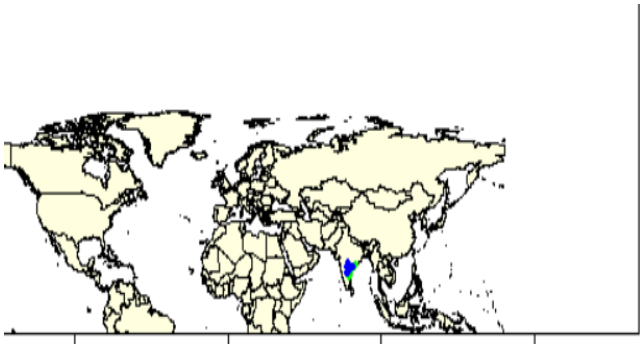
Sixteen features extracted from the data set are fed to the SVM and these are compared with other classifiers.

Image Grouping

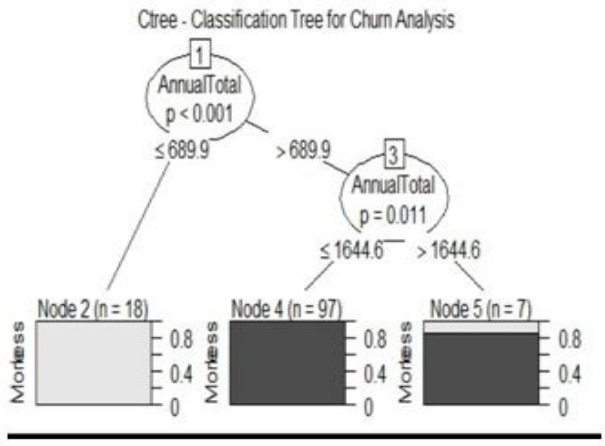
Classification output can be characterized by categorizing the same with a confusion matrix as demonstrated in Table-II.

Table II

Real group	Classification consequence	
	More	Less
More	TN	FP
Less	FN	TP



(a)



(b)

Figure 1a):Plot of Rainfall data with districts with more rainfall plotted with green and less with blue.

b) Decision Tree for Rainfall Data Classification.

Evaluation metrics

Evaluation of the classification models can be carried out using the metrics defined below[25-26]. The formulae for sensitivity, specificity, prevalence, and detection rate and detection prevalence are given in (10), (11), (11), (13) and (14)

$$\text{Sensitivity} = \frac{\text{TN}}{(\text{TN+FN})} \times 100 \quad (10)$$

$$\text{Specificity} = \frac{\text{TP}}{(\text{TP+FP})} \times 100 \quad (11)$$

$$\text{Prevalence} = \frac{\text{TN+FN}}{(\text{TP+FN+FP+TN})} \times 100 \quad (12)$$

$$\text{Detection rate} = \frac{\text{TN}}{(\text{TP+FN+FP+TN})} \text{x}100 \quad (13)$$

$$\text{Detection Prevalence} = \frac{\text{TN+FP}}{(\text{TP+FN+FP+TN})} \text{ X } 100 \quad (14)$$

Result Scrutiny

The performance of the proposed classifier SVM with linear kernel is examined and likened with SVM, Decision Tree, MLP and Naïve Bayes and the results are mentioned below.

Table III: Performance Measures

Classifier	Sensitivity	Specificity	Prevalence	Detection Rate	Detection Prevalence
SVM	100	100	18.75	12.5	12.5
Decision tree classifier	100	100	18.75	18.75	18.75
MLP Classifier	0	100	18.75	0	0
Naïve bayes classifier	83.3	84.62	18.75	15.62	28.12

Hence, it is identified from the results and findings with the comparative evaluations that SVM approach is better than other classifiers.

CONCLUSION

The research is related to the classification of rainfall data as belonging to various districts of Andhra Pradesh using SVM. The same is compared withdecision tree, MLP and Naïve Bayesmethods. The research has concluded that the SVM approach of classification is more effective and efficient.

REFERENCES

[1] Nadine Kashmar, MirnaAtieh, Ali Haidar, Identifying the Effective Parameters for Vertical Handover in Cellular Networks Using Data Mining Techniques, Procedia Computer Science, Volume 98, 2016, Pages 91-99, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.09.016.

[2] Ling Chen, Xue Li, Yi Yang, Hanna Kurniawati, Quan Z. Sheng, Hsiao-Yun Hu, Nicole Huang, Personal health indexing based on medical examinations: A data mining approach, Decision Support Systems, Volume 81, January 2016, Pages 54-65, ISSN 0167-9236, https://doi.org/10.1016/j.dss.2015.10.008.

[3] Eva Armengol, DionísBoixader, Francisco Grimaldo, Special Issue on Pattern Recognition Techniques in Data Mining, Pattern Recognition Letters, Volume 93, 1 July 2017, Pages 1-2, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2017.02.014.

[4] Gangin Lee, Unil Yun, HeungmoRyang, An uncertainty-based approach: Frequent itemset mining from uncertain data with different item importance, Knowledge-Based Systems, Volume 90, December 2015, Pages 239-256, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2015.08.018.

[5] TarekHamrouni, SarraSlimani, Faouzi Ben Charrada, A data mining correlated patterns-based periodic decentralized replication strategy for data grids, Journal of Systems and Software, Volume 110, December 2015, Pages 10-27, ISSN 0164-1212, https://doi.org/10.1016/j.jss.2015.08.019.

[6] Jia Wu, Shirui Pan, Xingquan Zhu, ZhihuaCai, Peng Zhang, Chengqi Zhang, Self-adaptive attribute weighting for Naive Bayes classification, Expert Systems with Applications, Volume 42, Issue 3, 15 February 2015, Pages 1487-1502, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2014.09.019.

[7] Liangxiao Jiang, Chaoqun Li, Shasha Wang, Lungan Zhang, Deep feature weighting for naive Bayes and its application to text classification, Engineering Applications of Artificial Intelligence, Volume 52, June 2016, Pages 26-39, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2016.02.002.

[8] P. Julian Benadit, F. Sagayaraj Francis, U. Muruganantham, Improving the Performance of a Proxy Cache Using Tree Augmented Naive Bayes Classifier, Procedia Computer Science, Volume 46, 2015, Pages 184-193, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.02.010.

[9] AleenaSwetapadma, AnamikaYadav, Protection of parallel transmission lines including inter-circuit faults using Naïve Bayes classifier, Alexandria Engineering Journal, Volume 55, Issue 2, June 2016, Pages 1411-1419, ISSN 1110-0168, https://doi.org/10.1016/j.aej.2016.03.029.

[10] ÖmerFarukArar, KürşatAyan, A Feature Dependent Naive Bayes Approach and Its Application to the Software Defect Prediction Problem, Applied Soft Computing, Available online 26 May 2017, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2017.05.043.

[11] ChuanChoong Yang, Chit Siang Soh, VooiVoon Yap, A non-intrusive appliance load monitoring for efficient energy consumption based on Naive Bayes classifier, Sustainable Computing: Informatics and Systems, Volume 14, June 2017, Pages 34-42, ISSN 2210-5379, https://doi.org/10.1016/j.suscom.2017.03.001.

[12] LokanathSarangi, Mihir Narayan Mohanty, SrikantaPattanayak, Design of MLP Based Model for Analysis of Patient Suffering from Influenza, Procedia Computer Science, Volume 92, 2016, Pages 396-403, ISSN1877-0509, https://doi.org/10.1016/j.procs.2016.07.396.

[13] JakubGajewski, David Vališ, The determination of combustion engine condition and reliability using oil analysis by MLP and RBF neural networks, Tribology International, Available online 23 June 2017, ISSN 0301-679X, https://doi.org/10.1016/j.triboint.2017.06.032.

[14] Ahmed F. Mashaly, A.A. Alazba, MLP and MLR models for instantaneous thermal efficiency prediction of solar still under hyper-arid environment, Computers and Electronics in Agriculture, Volume 122, March 2016, Pages 146-155, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2016.01.030.

[15] Yu Zhang, Shixing Wang, MLP technique based reinforcement learning control of discrete pure-feedback systems, Neurocomputing, Volume 168, 30 November 2015, Pages 401-407, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2015.05.087.

[16] TayyabWaqar, Mustafa Demetgul, Thermal analysis MLP neural network based fault diagnosis on worm gears, Measurement, Volume 86, May 2016, Pages 56-66, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2016.02.024.

[17] P.J. García Nieto, E. García-Gonzalo, J. Bové, G. Arbat, M. Duran-Ros, J. Puig-Bargués, Modeling pressure drop produced by different filtering media in microirrigation sand filters using the hybrid ABC-MARS-based approach, MLP neural network and M5 model tree, Computers and Electronics in Agriculture, Volume 139, 15 June 2017, Pages 65-74, ISSN 0168-1699, https://doi.org/10.1016/j.compag.2017.05.008.

[18] Dewan Md. Farid, Li Zhang, ChowdhuryMofizurRahman, M.A. Hossain, Rebecca Strachan, Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks, Expert Systems with Applications, Volume 41, Issue 4, Part

2, March 2014, Pages 1937-1946, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2013.08.089.

[19] Hamid Parvin, MiresmaeilMirnabiBaboli, Hamid Alinejad-Rokny, Proposing a classifier ensemble framework based on classifier selection and decision tree, Engineering Applications of Artificial Intelligence, Volume 37, January 2015, Pages 34-42, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2014.08.005.

[20] Om PrakashMahela, Abdul GafoorShaik, Recognition of power quality disturbances using S-transform based ruled decision tree and fuzzy C-means clustering classifiers, Applied Soft Computing, Volume 59, October 2017, Pages 243-257, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2017.05.061.

[21] P. Julian Benadit, F. Sagayaraj Francis, Improving the Performance of a Proxy Cache Using Very Fast Decision Tree Classifier, Procedia Computer Science, Volume 48, 2015, Pages 304-312, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2015.04.186.

[22] Shre Kumar Chatterjee, Saptarshi Das, Koushik Maharatna, Elisa Masi, Luisa Santopolo, IlariaColzi, Stefano Mancuso, Andrea Vitaletti, Comparison of decision tree based classification strategies to detect external chemical stimuli from raw and filtered plant electrical response, Sensors and Actuators B: Chemical, Volume 249, October 2017, Pages 278-295, ISSN 0925-4005, https://doi.org/10.1016/j.snb.2017.04.071.

[23] Aris Pagoropoulos, Anders H. Møller, Tim C. McAloone, Applying Multi-Class Support Vector Machines for performance assessment of shipping operations: The case of tanker vessels, Ocean Engineering, Volume 140, 1 August 2017, Pages 1-6, ISSN 0029-8018, https://doi.org/10.1016/j.oceaneng.2017.05.001.

[24] Abdulla Amin Aburomman, Mamun Bin IbneReaz, A novel weighted support vector machines multiclass classifier based on differential evolution for intrusion detection systems, Information Sciences, Volume 414, November 2017, Pages 225-246, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2017.06.007.

[25] Michael E. Cholette, Pietro Borghesani, Egidio Di Gialleonardo, Francesco Braghin, Using support vector machines for the computationally efficient identification of acceptable design parameters in computer-aided engineering applications, Expert Systems with Applications, Volume 81, 15 September 2017, Pages 39-52, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2017.03.050.

[26] Madson L. Dantas Dias, Ajalmar R. Rocha Neto, Training soft margin support vector machines by simulated annealing: A dual approach, Expert Systems with Applications, Volume 87, 30 November 2017, Pages 157-169, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2017.06.016.

[27] Sebastián Maldonado, Juan Pérez, Cristián Bravo, Cost-based feature selection for Support Vector Machines: An application in credit scoring, European Journal of Operational Research, Volume 261, Issue 2, 1 September 2017, Pages 656-665, ISSN 0377-2217, https://doi.org/10.1016/j.ejor.2017.02.037.

[28] O.P.Korobeynikov, and A.A.Trifilova."The Integration of the Strategic and Innovation Management", Management in Russia and Abroad, vol.4, 2001.URL:www.cfin.ru/press/management/.(27.11.2015).

[29] G.Y.Goldstein. "Strategic Innovation Management: Trends, Technology, and Practice: A Monograph". Taganrog: Publishing House TRTU, 2002.

[30] E.A.Gorbashko."Quality Management and Competitiveness". St. Petersburg: Publishing HouseSPb GUEF, 2008.

[31] I.Ansoff. "Strategic Management". Moscow: Ekonomika, 1998.

[32] P.Drucker. "Management Challenges in the 21stCentury". Moscow: Publishing House "Williams", 2003.

[33] P.Drucker, J.A.Makyarello. "Management". Moscow: Publishing House "Williams", 2010.

[34] M.H.Meskon, M.Albert and F.Hedouri. "Fundamentals of Management". Moscow: Delo, 1998.

[35] J.J.Lamben, I.Shulingand, R.Chumpitas."Market-Oriented Management. Textbook". St. Petersburg: Peter, Lider, 2010.

[36] P.N.Zavlin, A.K.Kazantsev and L.E.Mindeli. "Basics of Innovation Management: Theory and Practice". Moscow: "Unity", 2000.

[37] J.-J.Lamben. "Strategic Marketing. The European Perspective (Translated from French)". St. Petersburg: Nauka, 1996.

[38] E.V.Tolkacheva. "Strategic Controlling in the Enterprise Management System", Management in Russia and Abroad, vol.4, pp.109-118, 2004.

[39] S.M.Rezer. "Fundamentals of modeling optimal logical systems delivery. Transport Innovations", Scientific and Technical Journal, vol.3(18), 2014.