ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 6, Issue 10, October 2019

Classified and Misclassified data in Soft Set Environment

[1] Payal Chhabra, [2] Poonam
[1][2] Department of CSE, Rayat Bahra University

Abstract: To manage characterization for huge information, information or data filtering and cleansing are preferred as preprocessing steps. For the most part it evacuate noisy, errors and conflicted data and results misclassification. In this paper, we performed examination of misclassified data and recognize how much information is should be redressed to get important data. To exhibit this idea, we have utilized AirTrafficDataset from Statistical Computing Statistical Graphics to analyze misclassified content in informational index. Two directed classifiers are used: Support vector Machine and decision tree. The results shows that out of these classifiers, SVM classify 85% of the data correctly and only 15% of data has misclassification.

Index Terms – Classification, Machine Learning, soft set, Misclassification

1. INTRODUCTION

Right now buzz is around huge information. In information storm period, information is created from particular sources and gathered together inside the database. The need isn't to oust the information rather crowd it for later use. This outcomes in the development of this enormous information. Enormous information reflect conceivably gigantic data which is unmanageable through ordinary methods. The information falling under this classification requires unique instruments for the executives. Huge Data has extraordinary effect on the general public. Its rise keeps on pulling in different considerations as far as innovation. Informal organizations are incredible wellspring of Big information. Enormous information develop as aureate scurry for a social business. In 2010 information created over the world was around 1024 Exabyte and in 2014 around 7168 Exabyte a year [1]. It has been assessed that 2.5 EB information are creating each day [2,3] from our atmosphere information to online life information like transferring photographs (Instagram clients post around 2 lac photographs for each day [4]), recordings and other data. Utilization of gadgets, for example, advanced cells, PCs, PCs, sensors and tablets and so on is expanding. We are utilizing these gadgets and entering in the time of "Enormous Data". Enormous information is connected with 5Vs Volume, Velocity, Variety, Value and Veracity. Volume alludes to the measure of information that is being taken care of and used so as to get the ideal outcomes. Speed is about the information makes a trip starting with one point then onto the next because of high demands that end clients have for spilled information over various gadgets. Assortment speaks to various sort of information that is put away, researched and used. Worth is about the nature of information that is put away and its further utilization. Veracity manages consistency of enormous data.[5] These 5Vs offer ascent to gigantic complex information and to separate helpful data from existing huge delicate sets, information mining requires. While dealing with huge information on better places. mining turns into a challenge.[8,9,10]. The enormous measure of information in industry is of no utilization whenever contained unusable data. For extraction of sound information, named or unlabeled classifications are picked. While grouping huge information there can be some off-base arrangement that offer age to misclassification, it's anything but a bug, it is only an unsatisfactory order that ought to be amended and turns into a further challenge[6,7].

In this paper, we analyze a dataset to apply two classifier and identify misclassified data. This paper has five sections: First section discusses the flow and introduction of fog computing. Second section discusses the literature review and third section describes the various classification methods such as support vector machine and decision tree. Fourth section discusses the formulated problem. Fifth section describes the dataset and experimental evaluation. last section discusses the conclusion and its future scope.

2. LITERATURE REVIEW

In earlier years, there are a few examinations dependent on misclassification. For instance, In year 2004, Hout et.al[14] gathered the study of disease transmission informational index utilizing inactive class. In this paper examination of information is performed of randomized information utilizing log straight model and results misclassification. In year 1999, Brodley et. al[12] depicted a methodology for directed discovering that worked to discover and dispense with mislabeled information. A methodology is assessed on five datasets and exactness is thought about utilizing different procedures. In this, preparation set and testing set is considered as 0.9,0.1 and 40% misclassification is come about. In year 2009, Miranda, Garcia and Carvalho et. al[13] concentrated on bioinformatics dataset. As misclassification in natural information impacts the forecast presentation of classifier. The paper results high precision, in view of renaming the information in the wake of expelling misclassified information and perform mixture technique. In year 2005, Caudill et. al [11] chipped away at distinguishing and redressing incorrectly ordered information. This paper showed a 70% misclassification and applied a logit model on misclassified information which depends on logarithmic.

3. DIFFERENT CLASSIFICATION ALGORITHMS:

This section discusses the two classification methods and table1 contains the comparison analysis of SVM and decision tree based on different parameters.

1. Support Vector Machine:

SVM (Support Vector Machine) is introduced by both Cortes and Vapnik, is a technique that confines tests into two unique classes by drawing a hyper plane between them. Right when working with a various class plan issue, SVM gatherings tests into one of the two principal classes and further isolates each class until the decider class is acquired[15]. The primary objective of SVM is to make the Hyper plane and expand the edge, between the isolated positive examples and negative samples.SVM would then be able to anticipate the class of unlabeled examples through scrutinizing the side of independent plane. SVMs can deal with straight just as nonlinear arrangement issues. Utilizing this distinct and non-detachable assemblage is frequently dealt with by SVMs in the direct and nonlinear event[16].

2. Decision Tree:

Decision Tree can deal with enormous number of sources of info like as content, numbers and alphanumeric. This procedure manages systems can change with various bundles utilized or various stages utilized. Using the methodology, for example, fluffy principles or choice strategies, determination bush are utilized to deal with huge measure of information. Choice tree for the most part parts the information which thusly can be spared and further, the thought can be gathered again [17]. However, one weakness of this technique is that, on the off chance that there is any adjustment in information, at that point it could change the general consequences of information. Choice methodology technique can be utilized in therapeutic fields like hub is individual is male or female at that point further level is age, more noteworthy than 40 or not exactly and last experiencing any sickness or not. Various methodologies of choice tree are: C4.5, J48, CHAID, Iterative Dichotomiser 3(ID3) and CART (Classification And Regression Tree) etc[18].

Table1: comparison table of support vector machine and decision tree.

Classifier Name	Support Vector Machine	Decision Tree
Principle Based	Dimensionality of feature Space	Attribute value testing
Training Speed	Fast	Fast
Easy to interpret and understand	No	Yes
Prediction speed	Fast	Slow
Best to handle Dimensionality	High	High
Prediction Accuracy	High	Low

4. PROBLEM FORMULATION:

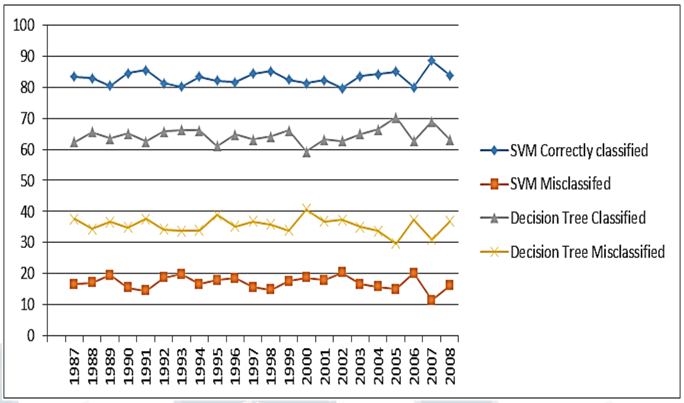
In software engineering condition, information is expanding step by step as delicate sets. To separate useable data from these informational indexes information mining procedure is utilized. In air transportation, informational collections are basic. Since it contains flight landing time, climate data, flight delays and some more. During information mining on this sort of information, order is performed to arrange the information like which flight is postponed and for what amount of time? In any case, that time issues can happen because of tremendous sum and diverse assortment of information and information cannot be right ordered. This issue turns into the test at that point wrong characterized information ought to distinguish and correct. So this paper depicts the misclassified information that implies how much information isn't all around arranged.

5. DATASET AND EXPERIMENTAL EVALUATION:

For this investigation informational index gathered of air traffic from 1987 to 2008 from Statistical Computing Stastical Graphics. This informational index incorporate 29 characteristics that are Year; Month from 1-12; DayofMonth 1-31; DayofWeek here 1 for monday, etc; DepTime is genuine takeoff time of flight; CSRDepTime is plan flight time; ArrTime is real entry time; CSRArrTime is plan landing time; UniqueCarrier is one of a kind specialist organization esteem; FlightNum is number of each flight; TailNum is plane seek after worth; ActualElaspedTime; AirTime; ArrDelay is entry delay; DepDelay is takeoff delay; Origin is code of takeoff place; Dest is spot code; Distance is to what extent; Taxiln is taxi in time; TaxiOut is taxi out time; Cancelled flight was dropped or not; CancellationCode A, B, C, D for example bearer, weather,NAS and security separately; Diverted for example 1 or 0 yes or no; CarrierDelay; WeatherDelay; NASDelay; SecurityDelay; LateAircraftDelay[19]. This dataset is characterized premise on one predefined classes. Results has been evaluated as shown in table 2.

Table2: Correctly classified and misclassified data from years 1987 to 2008

YEAR	Support vector machine		Decision Tree	
	Correctly Classified	Misclassified	Correctly Classified	Misclassified
1987	83.49	16.51	62.41	37.59
1988	82.98	17.02	65.57	34.43
1989	80.56	19.44	63.45	36.55
1990	84.56	15.44	65.14	34.86
1991	85.55	14.45	62.47	37.53
1992	81.32	18.68	65.74	34.26
1993	80.19	19.81	66.32	33.68
1994	83.45	16.55	66.14	33.86
1995	82.15	17.85	61.15	38.85
1996	81.58	18.42	64.84	35.16
1997	84.39	15.61	63.16	36.84
1998	85.23	14.77	64.17	35.83
1999	82.47	17.53	66.15	33.85
2000	81.36	18.64	59.32	40.68
2001	82.25	17.75	63.16	36.84
2002	79.67	20.33	62.73	37.27
2003	83.52	16.48	65.03	34.97
2004	84.30	15.70	66.34	33.66
2005	85.01	14.99	70.34	29.66
2006	79.98	20.02	62.68	37.32
2007	88.65	11.35	69.09	30.91
2008	83.91	16.09	63.23	36.77



Graph 1: Results of classified and misclassified data from 1987 to 2008

6. Conclusion:

This paper talks about enormous information ideas and demonstrates the materialness of conventional five methodologies on the specific dataset and inspires to address the misclassification for sometime later. In any case, initially how much information is mistakenly arranged ought to be assessed. This paper closed, in the wake of grouping air transportation information utilizing two classifiers (SVM, Selection Procedure) that Support vector machine performs better on a normal, it assessed 15% misclassification just and accurately arranged information is 85% information and in future this misclassification can be amended utilizing a few strategies like different ascription, revised score estimation and a lot all the more managing correction of enormous measure of information, to achieve 100% precision.

REFRENCES:

[1] Villars, Richard L., Carl W. Olofson, and Matthew Eastwood. "Big data: What it is and why you should care." White Paper, IDC (2011)

[2] Bello-Orgaz, Gema, Jason J. Jung, and David Camacho. "Social big data: Recent achievements and new challenges." Information Fusion 28 (2016): 45-59.

[3] IBM, Big Data and Analytics, URL http://www-01.ibm.com/software/data/bigdata/what-isbig-data.html (2015)

[4] Infographic, The Data Explosion in 2014 Minute by Minute, 2015. URL http://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic

[5] Tole, Alexandru Adrian. "Big data challenges." Database Syst J 4, no. 3 (2013): 31-40.

[6] Herzig, Kim, Sascha Just, and Andreas Zeller. "It's not a bug, it's a feature: how misclassification impacts bug prediction." In Proceedings of the 2013 International Conference on Software Engineering, pp. 392-401. IEEE Press, 2013.

[7] Kochhar, Pavneet Singh, Tien-Duy B. Le, and David Lo. "It's not a bug, it's a feature: does misclassification affect bug localization?." In Proceedings of the 11th Working Conference on Mining Software Repositories, pp. 296-299. ACM, 2014.

[8] Labrinidis, Alexandros, and Hosagrahar V. Jagadish. "Challenges and opportunities with big data." Proceedings of the VLDB Endowment 5, no. 12 (2012): 2032-2033.

[9] Wu, Xindong, Xingquan Zhu, Gong-Qing Wu, and Wei Ding. "Data mining with big data." ieee transactions on knowledge and data engineering 26, no. 1 (2014): 97-107.

[10] Fayyad, Usama M. "Data mining and knowledge discovery: Making sense out of data." IEEE Expert: Intelligent Systems and Their Applications 11, no. 5 (1996): 20-25.

[11] Caudill, Steven B., and Franklin G. Mixon. "Analysing misleading discrete responses: A logit model based on misclassified data." Oxford Bulletin of Economics and Statistics 67, no. 1 (2005): 105-113.

[12] Brodley, Carla E., and Mark A. Friedl. "Identifying mislabeled training data." Journal of Artificial Intelligence Research 11 (1999): 131-167.

[13] Miranda, André LB, Luís Paulo F. Garcia, André CPLF Carvalho, and Ana C. Lorena. "Use of classification algorithms in noise detection and elimination." In International Conference on Hybrid Artificial Intelligence Systems, pp. 417-424. Springer Berlin Heidelberg, 2009.

[14] Van den Hout, Ardo, and Peter GM Van der Heijden. "The analysis of multivariate misclassified data with special attention to randomized response data." Sociological Methods & Research 32, no. 3 (2004): 384-410.

[15] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20, no. 3 (1995): 273-297.

[16] O. Okun, G. Valentini, (Eds.), Supervised and Unsupervised Ensemble Methods and their Applications Studies in Computational Intelligence, vol. 126, Springer, Heidelberg, 2008.

[17] Lior Rokach and Oded Maimon,IEEE Transaction On System, Man and Cybernetics Part C, Vol 1, No. 11, November Top Down Induction Of Decision Tree Classifier-A Survey,2002

[18] Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." (2007): 3-24.

[19] Statistical Computing Statistical Graphics http://stat-computing.org/dataexpo/2009/the-data.html