Misclassification in Big Data Soft Set Environment

Jyoti Arora
Dept. Computer Engineering and Technology
Guru Nanak Dev University
Amritsar, India

Kamaljit Kaur
Dept. Computer Engineering and Technology
Guru Nanak Dev University
Amritsar, India

ABSTRACT
In order to deal with classification for large data, data filtering and data cleansing are used as preprocessing methods. Generally it remove noisy data, misclassified data, errors and inconsistent data and results unreliable classification. Because sometimes cleaned data can also affect the prediction accuracy or other testing. In this paper, we performed analysis of misclassified data and identify how much data has been wrong classified. For future aspect, This misclassified data is need to be rectified to get valuable information. To demonstrate this concept, we have used Air Traffic dataset from Statistical Computing Statistical Graphics (SCSG) to examine misclassified content in data set. Five supervised classifiers are used: Support vector Machine, decision procedure, k-nearest neighbor, random forest and logistic regression. The results shows that out of these classifiers, SVM classify 86% of the data correctly and only 14% of data has misclassification.

Keywords
Misclassification, Big Data, Classification

1. INTRODUCTION
Currently buzz is around big data. In data deluge era, data is generated from distinct sources and collected together within the database. The priority is not to expel the data rather hoard it for later use. This results in the formation of this large data. Big data reflect potentially huge information which is unmanageable through normal techniques. The data falling under this category requires special tools for management. Big Data has great impact on the society. Its emergence continues to attract diverse attentions in terms of technology. Social networks are great source of Big data. Big data emerge as aureate haste for a social business. In 2010 data generated over the world was about 1024 Exabyte and in 2014 about 7168 Exabyte a year [1]. It has been estimated that about 2.5 EB data are generating every day [2,3] from our climate data to social media data like uploading photos (Instagram users post about 2 lac photos per day [4]), videos and other information. Usage of devices such as smart phones, laptops, personal computers, sensors and tablets etc is increasing. We are using these devices and entering in the age of "Massive Data". Big data is related with 5Vs Volume, Velocity, Variety, Value and Veracity. Volume refers to the amount of data that is being handled and utilized in order to get the desired results. Velocity is all about the data travels from one point to another due to high requests that end users have for streamed data over numerous devices. Variety represents different kind of data that is stored, investigated and utilized. Value is all about the quality of data that is stored and the further use of it. Veracity deals with consistency of big data.[5] These 5Vs give rise to huge complex data and to extract useful information from existing large soft sets, data mining requires. While handling large data on different places mining becomes a challenge.[8,9,10]. The large amount of data in industry is of no use if contained unusable information. For extraction of healthy data, labeled or unlabeled categories are chosen. While classifying big data there can be some wrong classification that give generation to misclassification, it is not a bug, it is just a unsuitable categorization that should be rectified and becomes a further challenge[6,7].

1.1 Classification Associated With Huge Data
The classification involving sophisticated info is a difficult activity like to extract needle from a heap. But distinction associated with major information helps to set details and compare the differences and similarities. Categorize data into different classes can be performed in two ways like human driven methods and machine driven methods. In human driven methods, for classification, machines or any type of algorithms are not used like classification of plants. Each Category is characterized by humans. But in machine driven, for classification, machines are used and patterns are visualized to differentiate it. In computer science, here are further different ways to classify data: (a) Using supervised learning and (b) Unsupervised mastering (c) Semi-supervised discovering and (d) Reinforcement learning. In Supervised Learning, machine is given with a specimen advices and mapped with all the productions similar to naive bayes and many more. In Unsupervised Learning, machine is given with sample of methods but there is no desired productions like clustering. Reinforcement Learning deals with just how application brokers should get systems in an environment to optimize output for instance game theory. In semi-supervised learning, unlabeled data is used for training it is somehow works as supervised learning like self training.

In order to formulate, use of huge-scale data and extract useable information, engineers in machine learning are facing critical challenges. There are some tools of machine learning that are not able to perform well in personal computers because of memory and time. Due to this, there exists a requirement for evolution of scalable approaches for big data. There are basic two approaches used for machine learning in big data: Supervised system learning approaches and Unsupervised system learning approaches as shown in Fig1. In this paper, applicability of supervised learning on the data that deals with mainly 3Vs (Velocity, Volume, Variety) is evaluated. The classifiers discussed are as follows:

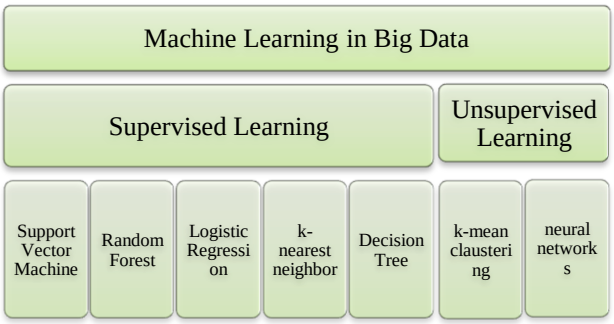


Fig1: Machine Learning approaches in big data

1.1.1 Supervised Learning Approaches

Due to some memory and time constraints, for considering large scale data sets, supervised learning is significant. In supervised learning approaches, machine is associated with specimen advices and it is mapped while using the production. Throughout this, input has some desired output. In this learning, each approach is based on training data. Here, using some dependent variables and some independent variables, correlation and usable information can be ectracted using supervision. Various approaches for supervised learning are used, some of them are: Support Vector Device, Selection Tree, Naive Bayes, Logistic Regression, k-nearest neighbor, and Random Forest etc.

1.1.1.1 Support Vector Machine

SVM (Support Vector Machine) is presented by both Cortes & Vapnik, is a procedure that isolates samples into two different classes by drawing a hyper plane between them. At the point when working with a numerous class arrangement issue, SVM groups samples into one of the two fundamental classes and further divides each class until the decider class is acquired[12]. The main goal of SVM is to make the Hyper plane and maximize the margin, between the separated positive samples and negative samples.SVM can then predict the class of unlabeled specimens through questioning the side of separate plane. SVMs are able to work on linear as well as nonlinear classification problems. Using this separable and non-separable compilation is often handled by SVMs in the linear and nonlinear event[21].

1.1.1.2 K-nearest neighbor

This approach is one of the preferred way of finding related objects. The nearest neighbor approach is based on calculation of common attributes of objects and form batches of them. These batches are then plotted to form cluster. Distinct clusters of similar attributes are formed which can easily be distinguished. In routine recognition, the k-Nearest Neighbors formula (k-NN) is often a non-parametric approach utilized for classification as well as regression. In both situations, feedback is comprised with k closest cases in feature space[22]. The particular productivity is determined by whether k-NN is usually used for explanation or even regression:

In k-nn classification, this result is a type membership. A thing is usually indexed by a majority votes of neighbors, with the thing remaining allotted to the category most common among its k nearby neighbors (k is a constructive integer) if k=1 then the thing is definitely allotted to the category of this single nearby neighbor.

In k-NN regression the average of K-nearest neighbor values has been considered as the outputs. In k-NN all the computation has been deferred and approximate locally until the classification. In k-NN regression, the actual outcome is definitely the value of thing. This importance is the standard of k neighbors. k-NN is known as lazy approach, instance based discovering . The k-NN formula is least complicated of all machine learning approaches.

For both these methods, it will be helpful to give excess weight for the donation of neighbors, so that the nearer neighbors add far as compared to the greater isolated such as weight of each neighbor is 1/ (neighbor's distance).

The value of N are taken from couple of things is actually the category (for k-NN classification) or the value of couple of things (for k-NN regression) is known. [11]

1.1.1.3 Random Forest

Random Forest is a gathering of decision trees. It is given by Breiman in 1999[23]. This classifier works on meta-learning which enhances the forecast quality by throwing votes among the trees and appointing the most voted class to the anticipated example. Random Forests are a standout machine learning strategy amongst the most capable, completely computerized. With no information readiness or demonstrating skill, experts can easily get surprisingly successful models. Random Forests is a fundamental part in the cutting edge information researcher's toolbox. Using number of decision trees, random forest can be constructed. Random forest can be applied for future prediction by using some continuous variables and it can work on probability also. Random Forests is a way that use numerous decision trees, sensible randomization, and outfit figuring out how to deliver amazingly precise prescient models, missing quality ascriptions, novel divisions, and laser-sharp giving an account of a record-by-record reason for profound information understanding[23]. In this each tress gives its own prediction then for combining its predictions average voting is used. Because of choosing the splitter randomly known as Random Forest.

1.1.1.4 Logistic Regression

Logistic Regression is one of the supervised machine learning that deals with binary values like 1 or 0. As we want to know, today rain will come or not like one will happen or other (1 or 0). Logistic regression is different from linear regression because logistic works on one aspect and it is affected by missing variables even though that variables are independent variables[24]. There ought to be no anomalies in the information, which can be surveyed by changing over the persistent indicators to institutionalized, or z-scores. There ought to be no high inter correlation among the indicators. This can be evaluated by a connection grid among the indicators.

1.1.1.5 Decision Tree

Decision Tree can handle large number of inputs like as text, numbers and alphanumeric. This strategy helps to deal with strategies can change with different packages used or different platforms used. Utilizing the procedures such as fuzzy rules or selection procedures, selection shrub are employed to handle large amount of data. Decision tree generally splits the data which in turn can be saved and further, the idea can be grouped again [25]. But one shortcoming of this strategy is that, if there is any change in data then it could possibly change the general results of data. Selection procedure strategy can be used in medical fields like node is person is male or female then further level is age, greater than 40 or less than and last suffering from any disease or not. Different approaches of decision tree are: C4.5, J48, CHAID, Iterative Dichotomiser 3(ID3) and CART (Classification And Regression Tree) etc[26].

1.1.2 Unsupervised Learning Approaches

Contrary to supervised learning, in unsupervised learning is performed on set of inputs. This learning can be performed to separate into different categories, without using any soft sets. In this learning, data is divided into different categories and basis on that categories similarities of the objects are checked without any supervision so known as unsupervised learning. Unsupervised learning doesn't bother about output so here no guide is present for guidance. Using this learning, no future prediction can be performed as labeled data and desired output is not considered. Different approaches of unsupervised system learning are clustering, anomaly detection, Independent component evaluation, face recognition and neural networks etc. like recognition of faces of different pet animals.

As there is not a particular rule to choose a classifier. Generally it depends on the number of features, type of dataset. As for first attempt naive bayes is recommended for classifying text and for numerical data neural network is preferred. But sometimes number of attributes are more than the samples taken then support vector machine classifier is to be chosen for better accuracy. At last, still it depends on problem like classification focused on computational speed or predictive accuracy etc. Table 1 describes the metrics based on that classifier can be chosen. For high prediction speed Support Vector Machine and for fast training speed Decision tree and Logistic Regression is to be preferred.

Table 1: Metrics of classifier

Classifier Name	Principle Based	Training Speed	Easy to interpret and understand	Prediction speed	Best to handle Dimensionality	Prediction Accuracy
Decision Tree	Attribute value testing	Fast	Yes	Slow	High	Low
Logistic Regression	Simple Hypothesis	Fast	Yes	Slow	High	High
Random Forest	Parallelization	Slow	Yes	Fast	High	High
Support Vector Machine	Dimensionality of feature Space	Fast	No	Fast	High	High
k-Nearest Neighbor	Distance Function (Memory Based)	Fast	Yes	Slow	High	High

In this paper, we analyze a big dataset to apply different classifiers and identify misclassified data. This paper describes five sections. The second section presents some related work as The Proof of Misclassification. The third area discusses about problem formulation. The next section discusses experiment evaluation and last section gives conclusions and suggests further challenges.

2. THE PROOF OF MISCLASSIFICATION

When two same instances are categorized into different classes that means misclassification. Generally, in classification problems, to evaluate the valuable information has been a challenge task. For performing categorization, it is very difficult to have proper function or equations. There is a requirement of approaches that could find relationship between input vectors and target vectors. The different learning methods are used for training and testing data. Therefore, to enhance accuracy of data training data should be improved. Whenever a category is undetectable then misclassification occurs. Whenever any category is usually hidden, or even unobservable, misclassification blunders usually are going to inevitable because it is not realizable to notice true class with accuracy. Even whenever a category is detectable, misconceptions can be built. Wrong classification can occur from usage of uncertain distinction techniques caused by realistic repression on information selection techniques. In Fig2 there is a confusion matrix that describes true positive, true negative and false positive ranges and example of these attributes is explained.

Example of Misclassification: Assume that problem is to classify patient data that is basis on person suffering from cold or not. There were 80 persons suffering from cold out of 120 but while classifying 90 persons are predicted as suffering from cold. This is how wrong classification occurs. So, 10 persons that are unsuitable at other category gives misclassification. Another explained classification of 165 students in which 105 students are below average and left are on average as shown in Fig3. So, True positive and true negative gives us best classification.

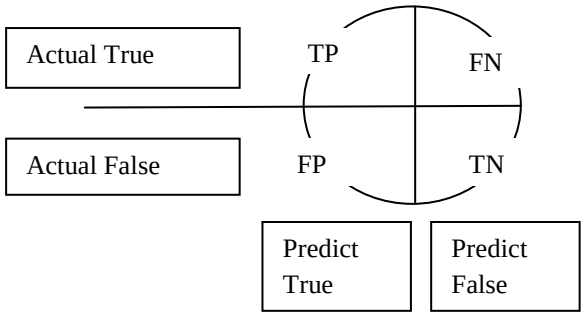


Figure2: Confusion Matrix

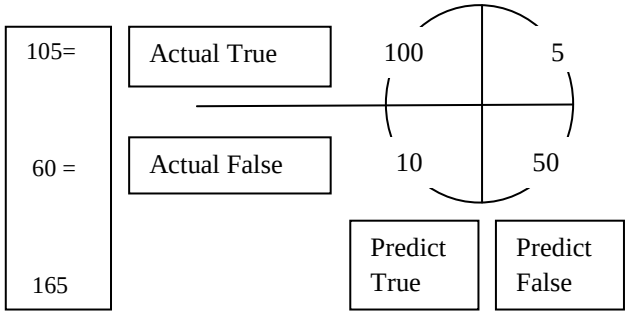


Figure3: Example of misclassification

From the above diagram, Misclassification can be calculated as ratio of bad prediction or misclassified data to total data.

Misclassification rate = (FP + FN) / (TP+TN+FP+FN)

While representing data in the form of sets then misclassification can be explained with example as: There are eight houses and categorization is based on which house is to purchase according to cost. And attributes considered are very cheap, cheap, costly and very costly. These eight houses are mapped with these parameters. As cheap houses are h1, h2, h3; very cheap house is h4; very costly house is h7; and costly houses are h5, h6. The values of these parameters collectively forms a bijective soft set. But if two parameters are considered out of these four then classification of houses are not in suitable manner. Therefore, misclassification occurs. There are some research papers that are reviewed for concept of misclassification.

2.1 Literature Survey

In previous years, there are several studies based on misclassification. For example, In year 2004, Hout et.al[16] collected epidemiology data set using latent class. In this paper analysis of data is performed of randomized data using log linear model and results misclassification. In year 1999, Brodley et. al[14] described a approach for supervised learning that worked to find out and eliminate mislabeled data. An approach is evaluated on five datasets and accuracy is compared using filtering techniques. In this, training set and testing set is considered as 0.9,0.1 and 40% misclassification is resulted. In year 2009, Miranda, Garcia and Carvalho et. al[15] focused on bioinformatics dataset. As misclassification in biological data effects the prediction performance of classifier. The paper results high accuracy, based on reclassifying the data after removing misclassified data and perform hybrid method. In year 2005, Caudill et. al [13] worked on identifying and rectifying wrong classified data. This paper indicated a 70% misclassification and applied a logit model on misclassified data which is based on logarithmic. In year 2009, Bilgic et. al[17] presented a RAC (Reflect and Correct) method that classify collectively and reflect the misclassification. To identify it, classifier is used that consisting features, samples further by using labels misclassification is corrected. In year 2005, Ciraco et. al[18] focused on improving the learning of classifier by chaining the cost ratio of misclassified data. This paper results that the false ratio or misclassification generally impacts maximum on training ratio and testing ratio of classifier. In year 2006, Finch et. al [30] defines the Misclassification can occur from the various statistical classification approaches. Each statistical approach is not fully accurate in practical. These approaches can be good as predictor. This paper results that when there is large difference of analyses to classify instances correctly that is known as misclassification. In year 2016, Ke Gong et. al [20]used bijective soft set to mine data from soft set environments. This paper proposed algorithm for finding misclassification degree and misclassified data. In this, proposed algorithm helped to discover fault data for analysis. Here, faulty data is defined but repairing of data is not considered.

3. PROBLEM FORMULATION

In computer Science environment, data is increasing day by day in the form of soft sets. To extract useable information from these data sets data mining process is used. In air transportation, data sets are very critical. Because it contains flight arrival time, weather information, flight delays and many more. During data mining on this type of data, classification is performed to categorize the data like which flight is delayed and for how much time? But that time problems can occur due to huge amount and different variety of data and data can be wrong classified. This problem becomes the challenge then wrong classified data should detect and rectify. So this paper describes the misclassified data that means how much data is not well categorized.

3.1 Dataset

For this experiment data set collected of air traffic from 1987 to 2008 from Statistical Computing Stastical Graphics. This data set include 29 attributes that are Year; Month from 1-12; DayofMonth 1-31; DayofWeek here 1 for monday and so on; DepTime is actual departure time of flight; CSRDepTime is schedule departure time; ArrTime is actual arrival time; CSRArrTime is schedule arrival time; UniqueCarrier is unique service provider value; FlightNum is number of each flight; TailNum is plane pursue value; ActualElaspedTime; AirTime; ArrDelay is arrival delay; DepDelay is departure delay; Origin is code of departure place; Dest is spot code; Distance is how long; Taxiln is taxi in time; TaxiOut is taxi out time; Cancelled flight was cancelled or not; CancellationCode A, B, C, D i.e. carrier, weather, NAS and security respectively; Diverted i.e. 1 or 0 yes or no; CarrierDelay; WeatherDelay; NASDelay; SecurityDelay; LateAircraftDelay[31]. This dataset is classified basis on one predefined classes.

4. EXPERIMENT EVALUATION

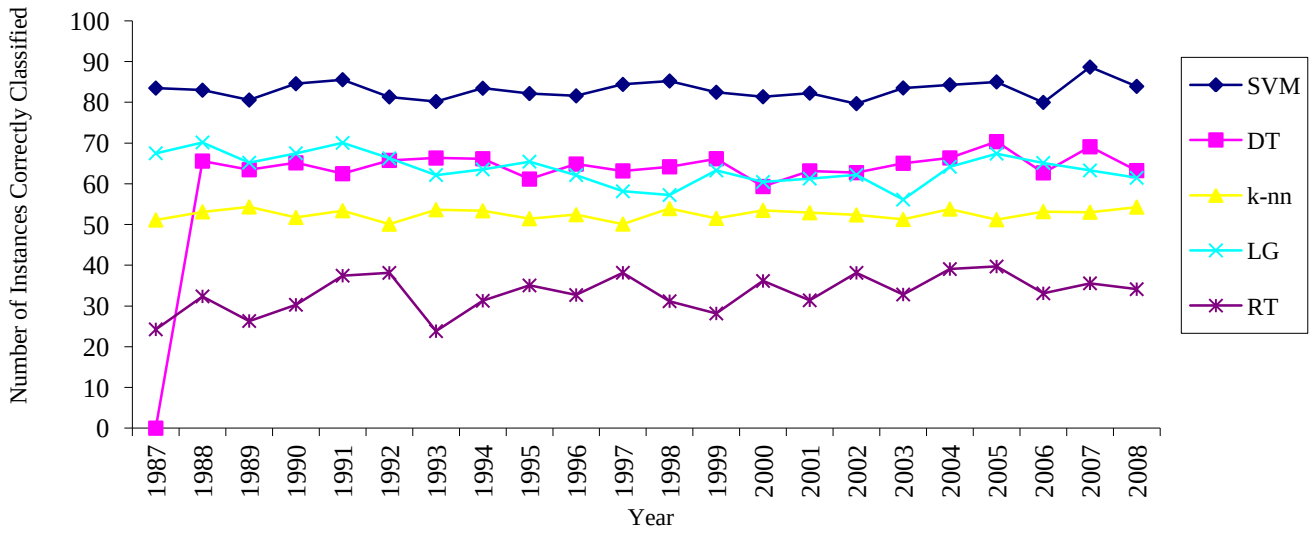
In this experiment, air traffic data is taken from 1987 to 2008 which has 29 parameters. This experiment is performed on jupyter big data tool using python 2.7. In this, five classifiers are used on each year data file to find out misclassified data. It is high volume data. Data set is divided into samples like (1,2000) (2000,4000) etc. Then classification is performed on the basis of arrival delay. Some features are selected that is origin of flight, destination of flight etc. Basis on that five classifiers (SVM, Decision Tree, Random Forest, k-nn, Logistic regression) are applied on each year data content separately to evaluate correctly and incorrectly classified as shown in Table2. Graph 1 represents correctly classified instances using five classifiers and Graph 2 represents incorrectly classified instances using these classifiers. This results that support vector machine gives less misclassified data. after that logistic regression and decision tree can be preferred.

5. CONCLUSION

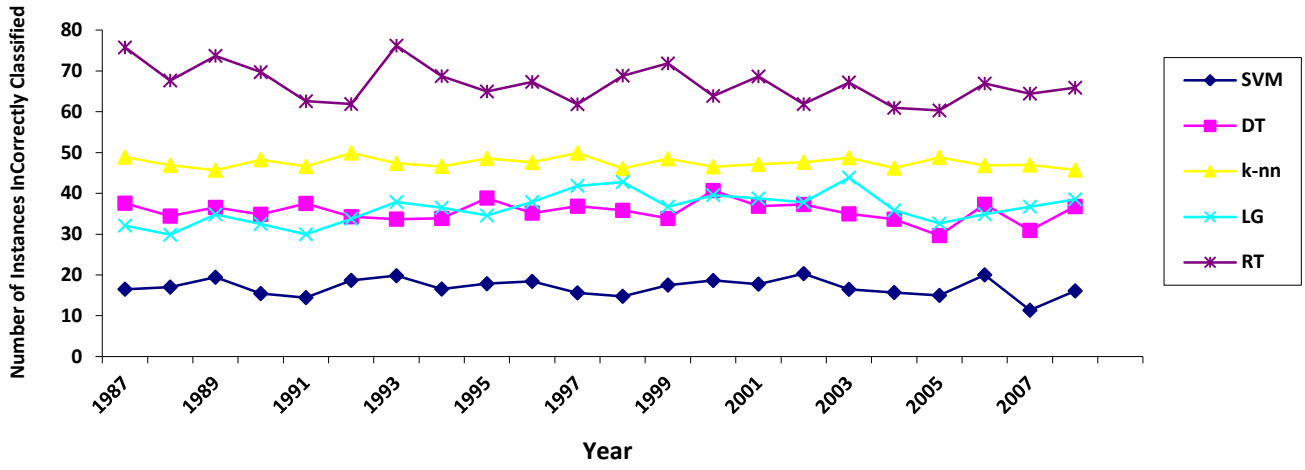
This paper discusses about big data concepts and shows the applicability of traditional five approaches on the particular dataset and motivates to correct the misclassification for future use. But firstly how much data is incorrectly classified should be evaluated. This paper concluded, after classifying air transportation data using five classifiers (SVM, Selection Procedure, Logistic Regression, k-nearest neighbor, Random Tree) that Support vector machine performs better on an average, it evaluated 14% misclassification only and correctly classified data is 86% data and in future this misclassification can be rectified using some techniques like multiple imputation, corrected score estimation and many more dealing with rectification of large amount of data, to attain 100% accuracy.

Table 2: Number of instances classified

YEAR	INSTANCES									
	Support Vector Machine		Decision Tree		Logistic Regression		Random Forest		k-nearest neighbor	
	Truely Classified	Mis-Classified	Truely Classified	Mis-Classified	Truely Classified	Mis-Classified	Truely Classified	Mis-Classified	Truely Classified	Mis-Classified
1987	83.49	16.51	62.41	37.59	67.90	32.10	24.25	75.75	51.11	48.89
1988	82.98	17.02	65.57	34.43	70.15	29.85	32.35	67.65	53.11	46.89
1989	80.56	19.44	63.45	36.55	65.17	34.83	26.31	73.69	54.32	45.68
1990	84.56	15.44	65.14	34.86	67.49	32.51	30.29	69.71	51.74	48.26
1991	85.55	14.45	62.47	37.53	70.03	29.97	37.43	62.57	53.39	46.61
1992	81.32	18.68	65.74	34.26	66.25	33.75	38.12	61.88	50.09	49.91
1993	80.19	19.81	66.32	33.68	62.14	37.86	23.81	76.19	52.63	47.37
1994	83.45	16.55	66.14	33.86	63.52	36.48	31.27	68.73	53.37	46.63
1995	82.15	17.85	61.15	38.85	65.43	34.57	35.04	64.96	51.43	48.57
1996	81.58	18.42	64.84	35.16	62.14	37.86	32.69	67.31	52.42	47.58
1997	84.39	15.61	63.16	36.84	58.15	41.85	38.15	61.85	50.12	49.88
1998	85.23	14.77	64.17	35.83	57.23	42.77	31.17	68.83	53.91	46.09
1999	82.47	17.53	66.15	33.85	63.26	36.74	28.16	71.84	51.52	48.48
2000	81.36	18.64	59.32	40.68	60.44	39.56	36.14	63.86	53.48	46.52
2001	82.25	17.75	63.16	36.84	61.26	38.74	31.39	68.64	52.89	47.11
2002	79.67	20.33	62.73	37.27	62.17	37.83	38.13	61.87	52.36	47.64
2003	83.52	16.48	65.03	34.97	56.12	43.88	32.82	67.18	51.25	48.75
2004	84.30	15.70	66.34	33.66	64.16	35.84	39.06	60.94	53.80	46.20
2005	85.01	14.99	70.34	29.66	67.39	32.61	39.68	60.32	51.18	48.82
2006	79.98	20.02	62.68	37.32	65.12	34.88	33.10	66.90	53.17	46.83
2007	88.65	11.35	69.09	30.91	63.28	36.72	35.56	64.44	53.03	46.97
2008	83.91	16.09	63.23	36.77	61.46	38.54	34.12	65.88	54.28	45.72



Graph1:Instances Correctly Classified in years 1987-2008 using five different classifiers



Graph2: Instances Incorrectly Classified in years 1987-2008 using five classifiers

6. REFERENCES

[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] Nodarakis, Nikolaos, Evaggelia Pitoura, Spyros Sioutas, Athanasios Tsakalidis, Dimitrios Tsoumakos, and Giannis Tzimas. "kdann+: A rapid aknn classifier for big data." In *Transactions on Large-Scale Data and Knowledge-Centered Systems XXIV*, pp. 139-168. Springer Berlin Heidelber 2016

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

[13] 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.

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

[15] 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.

[16] 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.

[17] Bilgic, Mustafa, and Lise Getoor. "Reflect and correct: A misclassification prediction approach to active inference." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 3, no. 4 (2009): 20.

[18] Ciraco, Michelle, Michael Rogalewski, and Gary Weiss. "Improving classifier utility by altering the misclassification cost ratio." In *Proceedings of the 1st international workshop on Utility-based data mining*, pp. 46-52. ACM, 2005.

[19] Nodarakis, Nikolaos, Evaggelia Pitoura, Spyros Sioutas, Athanasios Tsakalidis, Dimitrios Tsoumakos, and Giannis Tzimas. "kdann+: A rapid aknn classifier for big data." In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXIV*, pp. 139-168. Springer Berlin Heidelberg, 2016.

[20] Gong, Ke, Panpan Wang, and Yi Peng. "Fault-tolerant enhanced bijective soft set with applications." *Applied Soft Computing* (2016).

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

[22] Nodarakis, Nikolaos, Evaggelia Pitoura, Spyros Sioutas, Athanasios Tsakalidis, Dimitrios Tsoumakos, and Giannis Tzimas. "kdann+: A rapid aknn classifier for big data."In *Transactions on Large-Scale Data-and Knowledge-Centered Systems XXIV*, pp. 139-168. Springer Berlin Heidelberg, 2016.

[23] Breiman L: Random forests. Machine Learning 2001, 45:5-32.

[24] Mood, Carina. "Logistic regression: Why we cannot do what we think we can do, and what we can do about it." *European sociological review* 26, no. 1 (2010): 67-82.

[25] 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

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

[27] F. Salfner, M. Lenk and M. Malek, "A Survey of Online Prediction Methods," ACM Computing Surveys, vol. 22, no. 3, pp. 1-68, 2010.

[28] R. Jhawar, V. Piuri, and M. D. Santambrogio, "Fault tolerance management in IaaS clouds." In Satellite Telecommunications (ESTEL), 2012 IEEE 1st AESS European Conference, pp. 1-6, 2012.

[29] A. Avižienis, J.C. Laprie, B. Randell, and C. Landwehr. "Basic concepts and taxonomy of dependable and secure computing," IEEE Transactions on Dependable and Secure Computing, vol. 1, no. 1, pp. 11–33, 2004.

[30] Finch, W. H., & Schneider, M. K. (2006). Misclassification rates for four methods of group classification:Impact of predictor distribution, covariance inequality, effect size, sample size, and group size ratio. Educational and Psychological Measurement, 66, 240-257.

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