ISSN 2395-1621

A Survey: Image, Video & Text Categorization (K-NN, SVM, C5. 0)

[#1]Ms. Supriya Borse, [#2]Drs. Mrs. Neeta Deshpande

[1]supriyaborse123@gmail.com,
[2]deshpande_neeta@yahoo.com

[#1]M. E. II Computer Department,
[#2]Head of Computer Department

DYPCOE, Pune.

ABSTRACT

Categorization, technique of assigning data into different categories. Classification can be applied on different dataset such as video, audio, image and text. Different classification techniques are used on different types of dataset to achieve better performance. This paper focuses on survey of different classification algorithm such as SVM, k-NN, & C5.0. SVM has been widely applied on medical dataset, similarly k-NN & C5.0 algorithm are widely used in text categorization. Text categorization is the process of sorting and categorizing data into various types to find & retrieve specific information from given database within a time frame is one of the goal of data classification. According to current IT requirements, many classification algorithm are complex and sensitive to noise in database. C5.0 algorithm can be applied on large database to get optimum results. This can be applied on both discrete and continuous noise.

ARTICLE INFO

Article History

Received: 8th January 2017

Received in revised form :

8th January 2017

Accepted: 10th January 2017

Published online :

11th January 2017

Keywords: Data Mining, Classification, SVM, C5.0, K-NN.

I. INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. It is a technology used with great potential to help business and companies focus on the most important information of the data that they have to collect to find out their customer's behaviors. Intelligent methods are applied in order to extracting data pattern, by many stages like "data selection, cleaning, data integration, transformation and pattern extraction". Many methods are used for extraction data like "Classification, Regression, Clustering, Rule generation, Discovering, association Rule…etc. each has its own and different algorithms to attempt to fit a model to the data. This paper describes comparison about different classification methods based on their features [1].

Classification techniques in data mining are capable of processing a large amount of data. It can predict categorical class labels and classifies data based on training set and class label and hence can be used for classifying newly available data. In this paper classification method is considered, it focuses on a survey of various classification techniques that are most commonly used in data-mining.

Classification process is divided into two phases. The Training phase and the Testing phase. In training phase the pre-determined data and the associated class label are used for classification. The tuples used in training phase is called training tuples. In testing phase, the test data tuples are used to estimate the accuracy of classification rule. This paper represents comparison between three algorithms (Support Vector Machine, K-NN classifier and C5.0 algorithm) to show the strength and accuracy of each algorithm for classification in terms of performance efficiency and time complexity.

Steps for Classification:

1. Original dataset
2. Pre-Processing
3. Select the classifier
4. Classifier on training
5. Classifier on testing phase
6. Performance analysis of the selected classifier

Today people have access to enormous amount of videos, both on internet and on television. The amount of video that a viewer has to choose from is now so large that it is absurd

for human to go through it all to find video of their interest. Viewer will narrow their choices by looking for video within specific categories or genre [14]. Every day thousands of images are generated, which implies necessity to classify and access them by an easy and faster way. Image classification is one of the important and complex process in image processing. The main technique in which image can be classified efficiently is k-NN [15]. Text classification is an important part of text mining. Text classification process includes collection of data documents, data preprocessing, indexing and classification algorithm such as K-nearest neighbor classifier [2] [17], support vector machine [3] [16], C5.0 classifier [4].

K-NN is the nearest neighbor algorithm. The *k*-nearest neighbor's algorithm is a technique for classifying objects based on the next training data in the feature space. The algorithm operates on a set of d-dimensional vectors, $D = \{xi \mid i = 1. . . N\}$, where $xi \in kd$ denotes the i^{th} data point. The algorithm is initialized by selection *k* points in *k d* as the initial *k* cluster representatives or 'centroids' [2].

Support vector machines (SVM) are used to observe learning models with related learning algorithms. Algorithms are analyzing information to use for classification and reverting analysis. Set of training samples provided and every set is separately related with one of two classifications. A SVM training algorithm builds a model which assign new cases into one classification. This model makes it a non-probabilistic binary linear classifier [3].

C5.0 algorithm is the extension of C4.0 algorithm. It splits the user unrealized dataset by calculation of maximum information gain ratio. Every one subsample distinct by the primary split is after that split once more, frequently based on a dissimilar field, and the procedure replicate in anticipation of the subsamples cannot be split several additional [4].

II.RELATED WORKS

To give more prospective about the performance of the classification algorithm, this subsection describes and examines previous work done in field of data classification. The metrics taken into consideration are Learning Type, Speed, Accuracy, Scalability and Time complexity.

Rutvija et al. [5] studies about the time complexity and accuracy of C5.0 algorithm on a cloud network. This paper shows that C5 is a classifier which classifies the data in less time compare to other classifier. For generating decision tree the memory usage is minimum and it also improve the accuracy.

Amit et al. [6] have done the Comparative analysis of two algorithms; SVM and k-NN while considering certain parameters such as time complexity, improving efficiency and accuracy. These parameters are the major issue of concern in any classification algorithm. Experimental results show that the working of SVM algorithm is fast as compare to k-NN classifier.

Vanaja et al. [7] studies about the performance analysis of classification algorithm. The aim of this research paper is to study and discuss the various classification algorithms applied on different kinds of medical datasets and compares its performance. The classification algorithms with maximum accuracies on various kinds of medical datasets are taken for performance analysis. The result of the performance analysis shows the most frequently used algorithms on particular medical dataset and best classification algorithm to analyze the specific disease. This research paper also discusses the new features of C5.0 classification algorithm over C4.5 and performance of classification algorithm on high dimensional datasets.

III. DETAILED DESCRIPTION OF COMMON CLASSIFICATION ALGORITHM

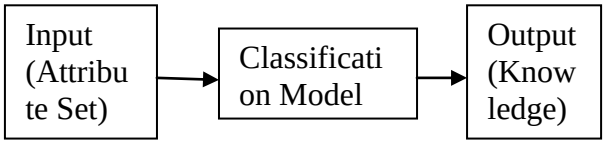

Fig. 2 Data mining using classification model [5].

3.1 K-NEAREST NEIGHBOR CLASSIFICATION

The k-nearest neighbor's algorithm is the nearest neighbor algorithm. It is a technique for classifying objects based on the next training data in the feature space. It is among simplest of all mechanism learning algorithms [2] [17].

- Input: Training Dataset
- Output: Class Labels

Algorithm:

1) Identify k-Nearest Neighbor (M)
   D= N * P
   PScenario $S^{1,\ldots,}S^{p}$.
   Scenario S^{i} Contain N features $S^{i} = \{ S_1^i, \ldots., S_N^i \}$
   Iterate loop M nearest neighbors times
   i) Collect S^{i} in the dataset
   ii) If q is not set or q < d (q. S^{i}): q ←d (q. S^{i}). t ← O^{i}
   iii) Iterate loop until i = p
   iv) Store q into vector c and t into vector.
2) Calculate mean output across 'r'

$$\bar{r} = \frac{1}{M} \sum_{i=1}^{M} ri$$

3) Return output value
4) Calculate distance between query instance and all training instance

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Parameter: -
M – Nearest Neighbors
N – Features collect in scenario

X_1, X_2 – two points in X-axis
Y_1, Y_2 – two points in Y-axis
i – Iteration number
D – Dataset matrix
P – Scenario matrix

k-nearest neighbor (k-NN) classification is an instance-based learning algorithm that has shown to be very effective in image classification [18]. k-NN classifier is best suited for classifying image due to its lesser execution time and better accuracy than other commonly used methods which include Hidden Markov Model and Kernel method. K-NN classifier has a faster execution time and is dominant than SVM [19]. To categorize an unknown document, the k-NN classifier ranks the document's neighbors among the training documents and uses the class labels of the k most similar neighbors. Similarity between two documents may be measured by the Euclidean distance [19]. The Euclidean distance is often chosen to determine the closeness between the data points in k-NN [22]. A distance is assigned between all pixels in a dataset. Distance is define as the Euclidean distance between two pixels, which is given by

$$d(x, y) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

3.2 SUPPORT VECTOR MACHINE

The SVM is a statistically robust learning method based on the *structural risk minimization* of the statistical learning theory [13]. The objective of the SVM is to find an optimal separating hyper plane which maximizes the margin between two classes of data in the kernel induced feature space [3].
Input: Training
Dataset Output: Class
Labels

Algorithm:

$$\sum_{i=1}^{m} \alpha i \; y^{(i)} \; K(x^{(i)}, x) + b$$

1) Kernel function (Radical basis function)

$$K(x, x') = \exp \frac{-(x - x')^2}{2\sigma}$$

2) Class y= -1, when output of scoring function is negative
3) Class y= 1, when output of scoring function is positive

Parameter:
X_i- i^{th} value of input vector
Y_{i-} i^{th} value of class label
a…i- is the coefficient associated with the $-i^{th}$ training dataset
b- Scalar value

Support Vector Machine has recently received much attention in the machine learning community [22]. Initially proposed as a binary classification method, SVM has not only been carefully motivated by statistical learning theory [23], but also been successfully applied to numerous domains, including object detection [24], handwritten digit recognition [25], and video categorization [3] [26]. Video classification is done based on feature. Feature consider for classification are color, shape, motion and other visual features. Color is an important attribute for image representation. Color histogram, which represents the color distribution in an image, is one of the most widely used color feature [26]. Color feature can be extracted using mean, variance and skewness.

3.3 C5 CLASSIFIER

The classifier is tested first to classify unseen data and for this purpose resulting decision tree is used. C5algorithm follows the rules of algorithm of C4.5 [5]. C5 algorithm has many features like:

- Large decision tree can be viewing as a set of rules which is easy to understand.
- C5 algorithm gives the knowledge on noise and missing data.
- Problem of over fitting & error pruning is solved by the C5 algorithm.
- In classification technique the C5 classifier can anticipate which attributes are relevant and which are not relevant in classification.

Input: Training Dataset
Output: Decision Tree
Algorithm:

1) Pick K attribute in dataset like $a_1, a_2, \ldots a_k$.
2) How frequently that combination occurs $a_1 = x_1, a_2 = x_2, \ldots a_k = x_k$.
3) Find the one with the highest information gain

$$\text{SplitInfo } A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{D} \right)$$

$$\text{Gain Ratio } (A) = \frac{Gain(A)}{SplitInfo(A)}$$

The split information value represent the potential information generated by splitting the training data set 'D' into 'v' partition, corresponding to 'v' outcomes an attribute A.

4) Create a decision node that split based on highest info gain.
5) Then apply recursion.

C5.0 algorithm is an improved version of C4.5 and ID3. This algorithm uses a value called Gain Ratio as a splitting criterion, unlike ID3algorithm where gain is used for splitting criteria in tree growth phase [19]. C5.0 is evolution of ID3 [27]. This algorithm handles both continuous and discrete attributes, in order to handle continuous attributes,

in C5.0 a threshold is created and then the list gets split into those whose attribute value is above the threshold and those that are less than are less than or equal to it [28].

This algorithm can deal with training data with attribute values by allowing an attribute value to be marked as.

3.3.1 IMPROVEMENT IN C4.5 FROM ID3 ALGORITHM

- C4.5 algorithm handles both continuous and discrete attributes.
- C4.5creates a threshold and makes the list of attributes having value above the threshold and less than or equal to the threshold.
- C4.5 algorithm also handles the training data with missing attributes values.
- In gain and entropy calculations the missing attribute values are not used.

3.3.2 IMPROVEMENT IN C5 FROM C4.5

ALGORITHM

- C5 is faster than C4.5
- Memory usage is more efficient in C5 than C4.5.
- C5 gets smaller decision trees in comparison with C4.5.
- The C5 rule sets have lower error rates on unseen cases.

IV.TABLE I (ADVANTAGES AND LIMITATIONS OF DIFFERENT CLASSIFICATIONALGORITHM)

Sr. no	Algorithm	Advantages	Limitations
1	K-nearest neighbor Algorithm	-Classes need not be linearly separable. -Zero cost of the learning process. -Sometimes it is robust with regard to noisy training data -Well suited for multimodal classes.	-Time to find the nearest neighbors in a large training set can be excessive. -It is sensitive to noisy or irrelevant attributes -Performance of algorithm depends on the number of dimensions used.
2	Support Vector Machine Algorithm	-High Accuracy -Work well even if data is not linearly separable in the base feature space.	-Speed & size requirement both in training & testing is more. -High complexity and extensive memory requirement for classification in many cases.
3	C5 Algorithm	-Build models can be easily interpreted. -Easy to implement -Can use both discrete and continuous noise -Deals with noise	-Small Variation in data can lead to different decision trees. - Does not work very well on a small training dataset. - Overfitting

V.TABLE II (FEATURE COMPARISON)

Sr.no	Feature	K-nn	SVM	C5.0
1	Learning type	Lazy learner	Eager learner	Eager learner
2	Accuracy	High-Robust	Significantly high	Good in many domains
3	Speed	Slow	Fast	Fast with active learning
4	Scalability	Efficient for small dataset	Efficient for uncertain dataset	Efficient for large dataset
5	Transparency	Rules	No rules (Black Box)	Rules
6	Missing value interpretation	Missing value	Sparse data	Missing value

VI.ACKNOWLEDGEMENT

The author would like to thank the publisher and researchers for making their resource available. We also thank the college authority for providing the required infrastructure and support. Finally we would like to extend our heartfelt gratitude to friends and family members.

VII.CONCLUSION

Classification is an important task in many data mining applications such as banking, medicine, scientific research and many more. Different classifiers are used for classifying different types of dataset. k-NN classifier is best suited for classifying image due to its lesser execution time and better accuracy than other commonly used methods . SVM is mainly used for video classification due its statistically robust learning method. C5.0 classifier is used for

classifying text dataset as it is decision tree algorithm which handles both continuous and discrete attributes.

REFERENCES

[1] AL-Nabi, DL Abd, and Shereen Shukri Ahmed. "Survey on classification algorithms for data mining: comparison and evaluation." International Journal of Computer Engineering and Intelligent Systems 4.8 (2013): 18-27.

[2] Samanthula, Bharath K., Yousef Elmehdwi, and Wei Jiang. "K-nearest neighbor classification over semantically secure encrypted relational data."IEEE transactions on Knowledge and data engineering 27.5 (2015): 1261-1273.

[3] Khorshed, Md Tanzim, ABM Shawkat Ali, and Saleh A. Wasimi. "A survey on gaps, threat remediation challenges and some thoughts for proactive attack detection in cloud computing." Future Generation computer systems28.6 (2012): 833-851.

[4] Vadivu, P. Senthil, and S. Nithya. "An Improved Privacy Preserving with RSA And C5. 0 Decision Tree Learning for Unrealized Datasets." International Journal 3.1 (2014).

[5] Pandya, Rutvija, and Jayati Pandya. "C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning." International Journal of Computer Applications 117.16 (2015).

[6] Sarode, Ashwini, and Mrs Saudagar Barde. "Secure Classification Of Encrypted Cloud Storage By Using SVM."

[7] Vanaja, S., and K. Rameshkumar. "Performance Analysis of Classification Algorithms on Medical Diagnoses-a Survey." Journal of Computer Science11.1 (2015): 31.

[8] Lin, Keng-Pei, and Ming-Syan Chen. "Privacy-preserving outsourcing support vector machines with random transformation." Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2010.

[9] Zhan, Justin Zhijun, LiWu Chang, and Stan Matwin. "Privacy Preserving K-nearest Neighbor Classification." IJ Network Security 1.1 (2005): 46-51.

[10] Galathiya, A. S., A. P. Ganatra, and C. K. Bhensdadia. "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning." International Journal of Computer Science and Information Technologies 3.2 (2012): 3427-3431.

[11] Ling-Li, L. I. "A Review on Classification Algorithms in Data Mining [J]."Journal of Chongqing Normal University (Natural Science) 4 (2011): 013.

[12] Vani, E., and S. Veena. "A Survey on Privacy Preserving k-NN Classification over Encrypted Data." International Journal of Engineering Science 6281 (2016).

[13] Joachims, Thorsten. "Text categorization with support vector machines: Learning with many relevant features." European conference on machine learning. Springer Berlin Heidelberg, 1998.

[14] Bhatt, Nirav. "A survey on video classification techniques."

[15] Kurian, Jipsa, and V. Karunakaran. "A survey on image classification methods."International Journal of Advanced Research in Electronics and Communication Engineering 1.4 (2012): pp-69.

[16] Pilászy, István. "Text categorization and support vector machines." The proceedings of the 6th international symposium of Hungarian researchers on computational intelligence. 2005.

[17] Joseph, Femi, and Nithin Ramakrishnan. "Text Categorization Using Improved K Nearest Neighbor Algorithm."

[18] Han, Eui-Hong Sam, George Karypis, and Vipin Kumar. "Text categorization using weight adjusted k-nearest neighbor classification." Pacific-asia conference on knowledge discovery and data mining. Springer Berlin Heidelberg, 2001.

[19] Rajeswari, K., et al. "Text Categorization Optimization By A Hybrid Approach Using Multiple Feature Selection And Feature Extraction Methods."

[20] Kaur, Manvjeet. "K-Nearest Neighbor Classification Approach for Face and Fingerprint at Feature Level Fusion." International Journal of Computer Applications 60.14 (2012).

[21] Turk, Matthew A., and Alex P. Pentland. "Face recognition using eigenfaces." Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on. IEEE, 1991.

[22] Lin, Wei-Hao, and Alexander Hauptmann. "News video classification using SVM-based multimodal classifiers and combination strategies." Proceedings of the tenth ACM international conference on Multimedia. ACM, 2002.

[23] Vapnik, Vladimir. The nature of statistical learning theory. Springer Science & Business Media, 2013.

[24] Papageorgiou, Constantine P., Michael Oren, and Tomaso Poggio. "A general framework for object detection." Computer vision, 1998. sixth international conference on. IEEE, 1998.

[25] Scholkopf, Bernhard, et al. "Comparing support vector machines with Gaussian kernels to radial basis function classifiers." IEEE transactions on Signal

Processing 45.11 (1997): 2758-2765.

[26] Suresh, Vakkalanka, et al. "Content-based video classification using support vector machines." International conference on neural information processing. Springer Berlin Heidelberg, 2004.

[27] Anyanwu, Matthew N., and Sajjan G. Shiva. "Comparative analysis of serial decision tree classification algorithms." International Journal of Computer Science and Security 3.3 (2009): 230-240.

[28] Quinlan, J. Ross. "Improved use of continuous attributes in C4. 5." Journal of artificial intelligence research 4 (1996): 77-90.

[29] Samanthula, Bharath K., Yousef Elmehdwi, and Wei Jiang. "K-nearest neighbor classification over semantically secure encrypted relational data."IEEE transactions on Knowledge and data engineering 27.5 (2015): 1261-1273.

[30] Lindell, Yehuda, and Benny Pinkas. "Privacy preserving data mining."Annual International Cryptology Conference. Springer Berlin Heidelberg, 2000.

[31] Gentry, Craig, and Shai Halevi. "Implementing Gentry's fully-homomorphic encryption scheme." Annual International Conference on the Theory and Applications of Cryptographic Techniques. Springer Berlin Heidelberg, 2011.

[32] Agrawal, Rakesh, and Ramakrishnan Srikant. "Privacy-preserving data mining." ACM Sigmod Record. Vol. 29. No. 2. ACM, 2000.

[33] Kantarcıoğlu, Murat, and Chris Clifton. "Privately computing a distributed k-nn classifier." European conference on principles of data mining and knowledge discovery. Springer Berlin Heidelberg, 2004.

[34] Xiong, Li, Subramanyam Chitti, and Ling Liu. "K nearest neighbor classification across multiple private databases." Proceedings of the 15th ACM international conference on Information and knowledge management. ACM, 2006.

[35] Bujlow, Tomasz, Tahir Riaz, and Jens Myrup Pedersen. "A method for classification of network traffic based on C5. 0 Machine Learning Algorithm." Computing, Networking and Communications (ICNC), 2012 International Conference on. IEEE, 2012.

[36] PANG, Su-lin, and Ji-zhang GONG. "C5. 0 classification algorithm and application on individual credit evaluation of banks." Systems Engineering-Theory & Practice 29.12 (2009): 94-104.

[37] Madzarov, Gjorgji, Dejan Gjorgjevikj, and Ivan Chorbev. "A multi-class SVM classifier utilizing binary decision tree." Informatica 33.2 (2009).

[38] Sugumaran, V., V. Muralidharan, and K. I. Ramachandran. "Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing." Mechanical systems and signal processing 21.2 (2007): 930-942

[39] Al-Harbi, S., et al. "Automatic Arabic text classification." (2008)..

[40] Lee, Kathy, et al. "Twitter trending topic classification." 2011 IEEE 11th International Conference on Data Mining Workshops. IEEE, 2011.

[41] Javidi, Mohammad Masoud, and Ebrahim Fazlizadeh Roshan. "Speech emotion recognition by using combinations of C5. 0, neural network (NN), and support vector machines (SVM) classification methods." J. Math. Comput. Sci 6 (2013): 191-200.

