International Journal of Research and Review
www.ijrrjournal.com E-ISSN: 2349-9788; P-ISSN: 2454-2237

Research Paper

Literature Survey on Data Classification Techniques

S.S.N. Alhady, M.A.N. Mohammad, A.A.A. Wahab, W.A.F.W. Othman

School of Electrical & Electronic Engineering, Universiti Sains Malaysia
14300 Nibong Tebal, Penang, Malaysia.

Corresponding Author: W.A.F.W. Othman

ABSTRACT

The growth of computer systems makes it easier to extract information from swarm of multiple nodes. These extracted big data will then be classified and divided into several categories for analysis. There are many ways to classify data, and this paper focuses on three, *i.e*. Support Vector Machine (SVM), K-Nearest Neighbor (kNN) and Fuzzy. The paper briefly explains each type of classifications and brings some of the physical systems that are using the data classification technique found on literatures. The goal is to summarize the existing approach towards data classification, guides the creation of new systems and point towards future directions.

Keywords: data classification, support vector machine, k-nearest neighbor, fuzzy

INTRODUCTION

Data classifications is broadly defined as a process of gathering, sorting and categorizing acquired data into relevant types, forms or any distinct class. Once the data have been classified, it may be used and protected more efficiently. The classification process of data will not only make the data to be easier to retrieved, but also easier to be located. Data classification is importance when we are dealing with data security, compliance and risk management.

Normally, any classification of data occurs in daily human activity. The data classification procedure for true classes has variously termed pattern recognition, discrimination, or supervised learning. Most of problem in daily life can be related as classification or decision using complex data. [1]

In this paper, we are focusing on three techniques which are Support Vector Machines (SVM), K-Nearest Neighbors (KNN) and Fuzzy.

SUPPORT VECTOR MACHINES (SVM)

Support vector machines are one of the methods of data classification with learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. [2]

Lin *et al*. [3] proposed a simple procedure which usually gives reasonable results. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one "target value" and several "attributes". The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes. The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes. [3]

Schölkopf [2] stated that Support-vector learning is very useful in two respects. First, it is quite satisfying from a theoretical point of view. SV learning is based on some beautifully simple ideas and provides a clear intuition of what learning

from examples is about. Second, it can lead to high performances in practical applications. It contains a large class of neural nets, radial basis function (RBF) nets, and polynomial classifiers as special cases. It is simple enough to be analyzed mathematically, because it can be shown to correspond to a linear method in a high-dimensional feature space nonlinearly related to input space. By the use of kernels, all necessary computations are performed directly in input space. This is the characteristic twist of SV methods. People are dealing with complex algorithms for nonlinear pattern recognition, regression or feature extraction, but for the sake of analysis people can pretend that they are working with a simple linear algorithm. [2]

Dumais [4] briefly described the results of experiments in which they use SVMs to classify newswire stories from Reuters. The goal of automatic text-categorization systems is to assign new items to one or more of a set of predefined categories on the basis of their textual content. Optimal categorization functions can be learned from labeled training examples. [4]

Osuna *et al*. introduced an SVM application for detecting vertically oriented and unclouded frontal views of human faces in gray-level images. For this discussion, face detection is an example of a natural and challenging problem for demonstrating and testing the potentials of SVMs. Many other real-world object classes and phenomena share similar characteristics. For example, tumor anomalies in MRI scans and structural defects in manufactured parts. A successful and general methodology for finding faces using SVMs should generalize well for other spatially well-defined pattern- and feature-detection problems. [5]

Platt from Microsoft Research [6] showed an SVM is a parameterized function whose functional form is defined before training. Training an SVM requires a labeled training set, because the SVM will fit the function from a set of examples. The training set consists of a set of *N* examples. Each example consists of an input vector, x_i, and a label, y_i, which describes whether the input vector is in a predefined category. There are *N* free parameters in an SVM trained with *N* examples. These parameters are called a_i. [6]

Mathieu *et al*. [7] from Department of Electrical and Computer Engineering, University of Iceland Hjardarhagi stated that Support Vector Machines (SVM) are used for the multisource classification rather than a Neural Network, which was used in their previous experiment with MP/EMP. The later added that the superiority of SVM, implementing structural risk minimization, over the neural classifiers, implementing empirical risk minimization. SVM aims to discriminate two classes by fitting an optimal separating hyper-plane to the training data within a multidimensional feature space, by using only the closest training samples. Thus, the approach only considers samples close to the class boundary and works well with small training set, even when high dimensional data sets are classified. [7]

Seyda *et al*. from The Pennsylvania State University, proposed an efficient SVM based active learning selection strategy which queries small pool of data at each iterative step instead of querying the entire dataset. The proposed method brings the advantage of efficient querying in search of the most informative instances, thus enabling active learning strategy to be applied to large datasets without high computational costs. [8]

Furey *et al*. from University of California developed a new method to analyses big data using support vector machines (SVMs). This analysis consists of both classification of the tissue samples, and an exploration of the data for miss-labeled tissue results. They demonstrated the method in detail on samples consisting of ovarian cancer tissues, normal ovarian tissues, and other normal tissues. The dataset consists of expression experiment results for 97,802 cDNAs for each tissue. As a result of computational analysis, a

tissue sample is discovered and confirmed to be wrongly labeled. Upon correction of this mistake and the removal of an outlier, perfect classification of tissues is achieved, but not with high confidence. They identify and analyze a subset of genes from the ovarian dataset whose expression is highly differentiated between the types of tissues. [9]

Camps-Valls *et al.* proposed the use of Support Vector Machines (SVM) for automatic hyperspectral data classification and knowledge discovery. In the first stage of the study, they used SVMs for crop classification and analyzed their performance in terms of efficiency and robustness, as compared to extensively used neural and fuzzy methods. Efficiency is assessed by evaluating accuracy and statistical differences in several scenes. [10]

Rossi and Villa investigated the use of Support Vector Machines (SVMs) for functional data analysis and focus on the problem of curves discrimination. SVMs are large margin classifier tools based on implicit nonlinear mappings of the considered data into high dimensional spaces thanks to kernels. They show how to define simple kernels that take into account the functional nature of the data and lead to consistent classification. [11]

Foody and Mathur, from University of Southampton evaluated the potential to target training data collection to regions that may contain useful training samples at the expense of those that will contribute insignificantly to classification by a SVM. The approach was based on the use of knowledge on the variables that influence the spectral response of the classes. [12]

Jayadeva *et al.* proposed a new nonparallel plane classifier, termed as the Twin Support Vector Machine (TWSVM) for binary data classification. TWSVMs also aim at generating two nonparallel planes such that each plane is closer to one of the two classes and is as far as possible from the other. TWSVMs slightly differ from SVMs in one fundamental way. In TWSVMs, they solve a pair of quadratic programming problems (QPPs), whereas, in SVMs, they solve a single QPP. In SVMs, the QPP has all data points in the constraints, but, in TWSVMs, they are distributed in the sense that patterns of one class give the constraints of the other QPP and vice versa. This strategy of solving two smaller sized QPPs, rather than one large QPP, makes TWSVMs work faster than standard SVMs. [13]

K-NEAREST NEIGHBORS (kNN)

K-Nearest Neighbors (kNN) algorithm is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. Data classified by a majority vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors. [14]

Li *et al*. demonstrated the applicability of the kNN method to SELDI proteomics data analysis. They showed by example that kNN is capable of finding a few ions capable of reliable discrimination between cancer and unaffected serum specimens using a published SELDI dataset. [15]

Guo *et al.* from School of Computing and Mathematics, [14] stressed that the k-Nearest-Neighbors (kNN) is a non-parametric classification method, which is simple but effective in many cases. For a data record *t* to be classified, its *k* nearest neighbors is retrieved, and this forms a neighborhood of *t*. In order for kNN to be less dependent on the choice of *k*, Guo proposed to look at multiple sets of nearest neighbors rather than just one set of kNN. the idea is to aggregate the support of multiple sets of nearest neighbors for various classes to give a more reliable support value, which better reveals the true class of *t*. [14]

Rosa and Ebecken from Universidade Federal do Rio de Janeiro, Brazil, presented a classification method based on the kNN-Fuzzy classification algorithm, supported by Genetic Algorithm.

It discussed on how to consider data clustering according to the Fuzzy logic and its consequences in the area of Data Mining. Analyses are made upon the results obtained in the classification of several data bases in order to demonstrate the proposed theory. [16]

Tran *et al.* introduced and developed a new density-based clustering algorithm, the so-called KNNCLUST. The proposed method is based on a combination of nonparametric kNN and kernel density estimation methods (kNN-kernel). It will be shown later in the text that kNN-kernel is not a good solution for estimating the "*true*" density of a distribution due to an overestimate of density in the tails of the distribution. [17]

KNNCLUST based on the combination of nonparametric *k*-Nearest-Neighbor and kernel (kNN-kernel) density estimation. The kNN-kernel density estimation technique makes it possible to model clusters of different densities in high-dimensional data sets. Moreover, the number of clusters is identified automatically by the algorithm. [17]

Jiang and Zhou from Nanjing University, [18] stated that, since kNN classifiers are sensitive to outliers and noise contained in the training data set, many approaches have been proposed to edit the training data so that the performance of the classifiers can be improved. Through detaching the two schemes adopted by the Depuration algorithm, two new editing approaches were derived. Moreover, they also have proposed that to use neural network ensemble to edit the training data for kNN classifiers. [18]

Yu *et al*. presented an efficient method, called *iDistance*, for k-Nearest Neighbor search in a high-dimensional space. *iDistance* partitions the data and selects a reference point for each partition. The data in each cluster are transformed into a single dimensional space based on their similarity with respect to a reference point. This allows the points to be indexed using a *B*+-tree structure and kNN search be performed using one dimensional range search. The choice of partition and reference point provides the *iDistance* technique with degrees of freedom most other techniques do not have. [19]

Zhang *et al.* of University of California [20] stated that the kNN classifier deals with the hugely multiclass nature of visual object recognition effortlessly. From a theoretical point of view, it has the remarkable property that under very mild conditions, the error rate of a kNN classifier tends to the Bayes optimal as the sample size tends to infinity. Instead of distorting the distance metric, they would like to bypass this cumbersome step and arrive at classification in one step. Here they propose to train a support vector machine (SVM) on the collection of nearest neighbors. This approach was well supported by ingredients in the practice of visual object recognition. [20]

Junno *et al.* introduced the use of the kNN method to identify different welding processes. Process information can be used to find suitable initialization parameters for welding machines or to predict the quality of welding spots using previously gathered data. In this study, the basic kNN method was found to be inadequate, and an extension to the kNN method was developed. The distance to the kNN was considered important information, and a similarity measure was formulated to provide this information to the user. According to the results, processes can be classified using the method and specific features. The similarity measure proved to be a valuable addition, which helps the user to decide whether the closest process is close enough to be classified as the same process. [21]

Nakai and Horton investigated the classification accuracy of three standard classification algorithms, namely the k-Nearest Neighbor's classifier, the binary decision tree, and the naive Bayes classifier; as well as the probabilistic model. To provide an additional baseline, they also compare the classification accuracy of using

kNN with the PAM120 local alignment distance instead of the expert identified features. In the first section they briefly describe the datasets, classifiers, and testing methodology. In the results section they give the accuracy results of cross-validation tests for the four methods and also for kNN with local alignment distances. Also in the results section, they investigate the effects of varying the *k* parameter and report which misclassifications are typical of the classifiers. [22]

Li *et al.* proposed combining a genetic algorithm (GA) and the kNN method to identify genes that jointly can discriminate between two types of samples (*e.g.* normal *vs.* tumor). First, many such subsets of differentially expressed genes are obtained independently using the GA. Then, the overall frequency with which genes were selected is used to deduce the relative importance of genes for sample classification. Sample heterogeneity is accommodated; that is, the method should be robust against the existence of distinct subtypes. They applied GA/KNN to expression data from normal versus tumor tissue from human colon. Two distinct clusters were observed when the 50 most frequently selected genes were used to classify all of the samples in the data sets studied and the majority of samples were classified correctly. Identification of a set of differentially expressed genes could aid in tumor diagnosis and could also serve to identify disease subtypes that may benefit from distinct clinical approaches to treatment. [23]

Kolahdouzan and Shahabi from University of Southern California, proposed a novel approach that by reducing the problem of distance computation in a very large network, in to the problem of distance computation in a number of much smaller networks plus some additional table lookups. The main idea behind their approach was termed Voronoi based Network Nearest Neighbor (VN3), is to first partition a large network in to smaller/more manageable regions. [24]

Seidl and Kriegel from Germany presented an adapted version of the multi-step algorithm for k-nearest neighbor search of medical imaging. The query object is denoted by *q*, and the parameter *k* specifies the requested number of neighbors. The basic structure of the algorithm is that it proceeds in two stages: In the first stage, a *k*-nearest neighbor search on the index is performed returning the *k* closest objects with respect to the filter distance function. For these *k* objects, the maximum d_{max} of the exact object distances is determined. In the second stage, a range query on the index is performed returning all objects that have a filter distance of at most d_{max}. For all of these candidates, the exact object distance is evaluated, and the *k* closest objects are reported. [25]

FUZZY

Fuzzy classification is the process of grouping element into fuzzy set whose membership function is defined by the truth value of a fuzzy propositional function. [26]

Abe *et al.* proposed a new method for extracting fuzzy rules directly from numerical input-output data for pattern classification. Fuzzy rules with variable fuzzy regions are defined by activation hyper-boxes which show the existence region of data for a class and inhibition hyper-boxes which inhibit the existence of data for that class. These rules are extracted from numerical data by recursively resolving overlaps between two classes. Then, optimal input variables for the rules are determined using the number of extracted rules as a criterion. [27]

Ishibuchi *et al.* from Osaka Prefecture University, introduced a genetic-algorithm-based method for selecting a small number of significant fuzzy if-then rules to construct a compact fuzzy classification system with high classification power. The rule selection problem was formulated as a combinatorial optimization problem with two objectives: to maximize the number of correctly classified patterns and to minimize the number of fuzzy if-then

rules. Genetic algorithms are applied to this problem. A set of fuzzy if-then rules was coded into a string and treated as an individual in genetic algorithms. The fitness of each individual is specified by the two objectives in the combinatorial optimization problem. [26]

Acharya *et al.* suggested the classification of certain diseases using artificial neural network (ANN) and fuzzy equivalence relations. The heart rate variability is used as the base signal from which certain parameters are extracted and presented to the ANN for classification. The same data is also used for fuzzy equivalence classifier. The feed forward architecture ANN classifier is seen to be correct in about 85% of the test cases, and the fuzzy classifier yields correct classification in over 90% of the cases. [28]

Zhong *et al.* proposed a method of automatic fuzzy clustering based on adaptive multi-objective differential evolution (AFCMDE), to perform the task of clustering for remote sensing imagery. In AFCMDE, the automatic fuzzy clustering problem is transformed into a multi-objective optimization problem. The values of J_m (weighted within-groups sum of squared errors as the objective function) and the X_B index are selected as the optimization functions to synchronously measure the clustering validity. To solve the multi-objective optimization problem, the multi-objective differential evolution (MODE) algorithm, as a new competitive optimized method was utilized. Some known advantages of differential evolution (DE) are, ease of use, robust, and having a powerful search capabilities. [29]

Burrough, *et al.* from Canada stated that, methods of fuzzy k-means have been used by other workers to overcome the problem of class overlap but their usefulness may be reduced when data sets are large and when the data include artifacts introduced by the derivation of landform attributes from gridded digital elevation models. Their paper presented ways to overcome these limitations using spatial sampling methods, statistical modeling of the derived stream topology, and fuzzy k-means using the *Distance* metric. Using data from Alberta, Canada, and the French pre-Alps it was shown how the proposed methods may easily create meaningful, spatially coherent landform classes from high resolution gridded differential evolution methods. [30]

González Rodríguez *et al.* utilized the use of the fuzzy scale of measurement to describe an important number of observations from real-life attributes or variables are first explored. In contrast to other well-known scales (like nominal or ordinal), a wide class of statistical measures and techniques can be properly applied to analyze fuzzy data. This fact is connected with the possibility of identifying the scale with a special subset of a functional Hilbert space. The identification can be used to develop methods for the statistical analysis of fuzzy data by considering techniques in functional data analysis and vice versa. In this respect, an approach to the ANOVA test was presented. It is later particularized to deal with fuzzy data. The proposed approaches are illustrated by means of a real-life case study. [31]

DISCUSSION

We have seen about thirty data classification works in the previous sections. Amongst the three techniques *i.e*. Support Vector Machines, K-Nearest Neighbors and Fuzzy, we have chosen kNN to be used in our future system, which is crowd sensor using multiple types of sensors. This is because kNN has several advantages like simplicity and effectiveness. It also robust to noisy data and more effective if the quantity of data is large.

ACKNOWLEDGEMENTS

Authors declare no conflict of interest. The authors also would like to thank Universiti Sains Malaysia for supported the work by Fundamental Research Grant Scheme (FRGS): Ministry of Education Malaysia (Grant number: USM/PELECT/6071239).

REFERENCES

1. D. Michie, D.J. Spiegelhalter, and C.C. Taylor. "Machine learning, neural and statistical classification." pp. 2-5. 1994.
2. B. Scholköpf. "SVMs-a practical consequence of learning theory." IEEE Intelligent Systems, pp 18-21. 1995.
3. C.J. Lin, C. W. Hsu, and C. C. Chang. "A practical guide to support vector classification." National Taiwan U., Available from: www.csie.ntu.edu. tw/cjlin/papers/guide/guide.pdf 2010
4. S. Dumains. "Using SVMs for text categorization." IEEE Intelligent Systems 13, vol. 4, pp 21-23. 1998.
5. E. Osuna, R. Freund, and F. Girosi. "Training support vector machines: an application to face detection." In Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on, pp. 130-136. IEEE, 1997.
6. J. Platt, "How to implement SVM" IEEE Intelligent Systems, pp 26-28. 1998
7. M. Fauvel, J.A. Benediktsson, J. Chanussot, and J. R. Sveinsson. "Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles." Geoscience and Remote Sensing, IEEE Transactions on 46, vol. 11, pp. 3804-3814. 2008.
8. S. Ertekin, J. Huang, L. Bottou, and C.L. Giles. "Learning on the border: active learning in imbalanced data classification." In Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, pp. 127-136. 2007
9. T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, and D. Haussler. "Support vector machine classification and validation of cancer tissue samples using microarray expression data." Bioinformatics 16, vol. 10, pp. 906-914. 2000.
10. G. Camps-Valls, L. Gómez-Chova, J. Calpe-Maravilla, J. D. Martín-Guerrero, E. Soria-Olivas, L. Alonso-Chordá, and J. Moreno. "Robust support vector method for hyperspectral data classification and knowledge discovery." Geoscience and Remote Sensing, IEEE Transactions on 42, vol. 7, pp. 1530-1542. 2004.
11. F. Rossi and N. Villa. "Support vector machine for functional data classification." Neurocomputing 69, vol. 7, pp. 730-742. 2006
12. G.M. Foody, and A. Mathur. "Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification." Remote Sensing of Environment 93, vol. 1, pp 107-117. 2004.
13. Jayadeva, R. Khemchandani, and S. Chandra. "Twin support vector machines for pattern classification." Pattern Analysis and Machine Intelligence, IEEE Transactions on 29, vol. 5, pp 905-910. 2007.
14. G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer. "KNN model-based approach in classification." In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, pp. 986-996. Springer Berlin Heidelberg, 2003.
15. L. Li, D. M. Umbach, P. Terry, and J. A. Taylor. "Application of the GA/KNN method to SELDI proteomics data." Bioinformatics 20, vol. 10, pp. 1638-1640. 2004.
16. J.L.A. Rosa and N.F.F. Ebecken. "Data mining for data classification based on the KNN-fuzzy method supported by genetic algorithm." In High Performance Computing for Computational Science—VECPAR 2002, pp. 126-133. Springer Berlin Heidelberg, 2003.
17. T. N. Tran, R. Wehrens, and L. Buydens. "KNN-kernel density-based clustering for high-dimensional multivariate data." Computational Statistics & Data Analysis 51, vol. 2, pp. 513-525. 2006.
18. Y. Jiang, and ZH. Zhou. "Editing training data for kNN classifiers with neural network ensemble." In Advances in Neural Networks–ISNN 2004, pp. 356-361. Springer Berlin Heidelberg, 2004.
19. C. Yu, B.C. Ooi, KL. Tan, and H.V. Jagadish. "Indexing the distance: An efficient method to knn processing." In VLDB, vol. 1, pp. 421-430. 2001.
20. H. Zhang, A.C. Berg, M. Maire, and J. Malik. "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition." In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, vol. 2, pp. 2126-2136. 2006.
21. H. Junno, P. Laurinen, E. Haapalainen, L. Tuovinen, and J. Röning. "Resistance spot welding process identification using an extended knn method." In Proc. IEEE

International Symposium on Industrial Electronics, vol. 1, pp. 7-12. 2005.
22. P. Horton, and K. Nakai. "Better Prediction of Protein Cellular Localization Sites with the k Nearest Neighbors Classifier." In Proc Int Conf Intell Syst Mol Biol., vol. 5, pp. 147-152. 1997.
23. L. Li, T.A. Darden, C.R. Weingberg, A.J. Levine, and L.G. Pedersen. "Gene assessment and sample classification for gene expression data using a genetic algorithm/k-nearest neighbor method." Combinatorial chemistry & high throughput screening 4, vol. 8, pp. 727-739. 2001.
24. M. Kolahdouzan, and C. Shahabi. "Voronoi-based k nearest neighbor search for spatial network databases." In Proceedings of the Thirtieth international conference on Very large data bases-Volume 30, pp. 840-851. VLDB Endowment, 2004.
25. T. Seidl, and HP. Kriegel. "Optimal multi-step k-nearest neighbor search." In ACM SIGMOD Record, vol. 27, vol. 2, pp. 154-165. ACM, 1998.
26. H. Ishibuchi, K. Nozaki, N. Yamamoto, and H. Tanaka. "Selecting fuzzy if-then rules for classification problems using genetic algorithms." Fuzzy Systems, IEEE Transactions on 3, vol. 3, pp. 260-270. 1995.
27. S. Abe, and MS. Lan. "A method for fuzzy rules extraction directly from numerical data and its application to pattern classification." Fuzzy Systems, IEEE Transactions on 3, vol. 1, pp. 18-28. 1995.
28. U.R. Acharya, P.S. Bhat, S.S. Iyengar, A. Rao, and S. Dua. "Classification of heart rate data using artificial neural network and fuzzy equivalence relation." Pattern Recognition 36, vol. 1, pp. 61-68. 2003.
29. Y. Zhong, S. Zhang, and L. Zhang. "Automatic fuzzy clustering based on adaptive multi-objective differential evolution for remote sensing imagery." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6, vol. 99, pp. 1-12. 2013.
30. P.A. Burrough, P.F.M. van Gaans, and R.A. MacMillan. "High-resolution landform classification using fuzzy k-means." Fuzzy sets and systems, vol 113, 1, 37-52. 2000.
31. G. González-Rodríguez, A. Colubi, and M.Á. Gil. "Fuzzy data treated as functional data: A one-way ANOVA test approach." Computational Statistics & Data Analysis 56, vol. 4, pp. 943-955. 2012.

How to cite this article: Alhady SSN, Mohammad MAN, Wahab AAA et.al. Literature Survey on Data Classification Techniques. International Journal of Research and Review. 2018; 5(10):214-221.

******