2009 Ninth International Conference on Intelligent Systems Design and Applications

# Evaluation Measures for Ordinal Regression

Stefano Baccianella, Andrea Esuli and Fabrizio Sebastiani
*Istituto di Scienza e Tecnologia dell'Informazione*
*Consiglio Nazionale delle Ricerche*
*Via Giuseppe Moruzzi 1 – 56124 Pisa, Italy*
{firstname.lastname}`@isti.cnr.it`

***Abstract*—*Ordinal regression* (OR – also known as *ordinal classification*) has received increasing attention in recent times, due to its importance in IR applications such as learning to rank and product review rating. However, research has not paid attention to the fact that typical applications of OR often involve datasets that are highly imbalanced. An imbalanced dataset has the consequence that, when testing a system with an evaluation measure conceived for balanced datasets, a trivial system assigning all items to a single class (typically, the majority class) may even outperform genuinely engineered systems. Moreover, if this evaluation measure is used for parameter optimization, a parameter choice may result that makes the system behave very much like a trivial system. In order to avoid this, evaluation measures that can handle imbalance must be used. We propose a simple way to turn standard measures for OR into ones robust to imbalance. We also show that, once used on balanced datasets, the two versions of each measure coincide, and therefore argue that our measures should become the standard choice for OR.**

***Keywords*-Ordinal regression; Ordinal classification; Evaluation measures; Class imbalance; Product reviews**

## I. Introduction

The problem of rating objects with values ranging on an ordinal scale is called *ordinal regression* (OR – also known as *ordinal classification*). OR consists of estimating a *target function* $\Phi : X \rightarrow Y$ which maps each object $x_i \in X$ into exactly one of an ordered sequence $Y = \langle y_1 \prec \ldots \prec y_n \rangle$ of *classes* (also known as "scores", or "ranks", or "labels"), by means of a function $\hat{\Phi}$ called the *classifier*. This problem lies in-between *single-label classification*, in which $Y$ is instead an unordered set, and *(metric) regression*, in which $Y$ is instead a continuous, totally ordered set (typically: the set $\mathbb{R}$ of the reals).

In recent years OR has witnessed an increased interest in the information retrieval (IR) community. One of the reasons is the fact that OR is one of the most important approaches to learning to rank (see e.g., [1], [2], [3]). The second reason is that OR is a natural choice for rating product reviews, a problem which has received increased attention within sentiment analysis and opinion mining (see e.g., [4]). This latter task (to which we will mostly refer in this paper) consists in attributing a score of satisfaction to consumer reviews of a product based on their textual content; this is akin to guessing, based on an analysis of the textual content of the review, the score the reviewer herself would attribute to the product. This problem arises from the fact that, while some online product reviews consist of a textual evaluation of the product *and* a score expressed on some ordered scale of values, many other reviews contain a textual evaluation only. These latter reviews are difficult for an automated system to manage, and associating them with a score in an automatic way would make them more manageable. While some researchers have used binary scores (i.e., classifying the reviews as Positive or Negative – see e.g., [5], [6], [7], [8]) or ternary scores (also including a Neutral class – see e.g., [9], [10]), others have tackled the more complex problems of attributing scores from an ordinal scale containing an arbitrary (finite) number of values (see e.g., [11], [12], [13], [14]). This scale may be in the form either of an ordered set of *numerical* values (e.g., 1 stars to 5 stars), or of an ordered set of *non-numerical* values (e.g., Poor, Fair, Good, Very Good, Excellent). The only difference between these two cases is that, while in the former case the distances between consecutive values are known, this is not true in the latter case.

## II. Imbalanced datasets and trivial classifiers

Despite this renewed interest in OR, research has seemingly not paid attention to the fact that the datasets OR tackles are often severely *imbalanced*, i.e., some classes are far more frequent than others. For example, the TripAdvisor-15763 hotel review dataset we have presented in [11], consisting of all the English-language reviews of hotels in Pisa and Rome available from the TripAdvisor[1] Web site at the beginning of May 2008, is severely imbalanced, since 45% of all the reviews have 5 stars, 34.5% have 4 stars, 9.4% have 3 stars, 7.2% have 2 stars, and only 3.9% have 1 star. This example is not isolated: in 2006 Jindal and Liu [15] crawled a corpus of 5.8 million reviews from Amazon[2], and found that, again on a scale of 1 star to 5 stars, 57.5% of the reviews had 5 stars, 20.0% had 4 stars, 8.7% had 3 stars, 5.5% had 2 stars, and 8.3% 1 star. The fact that online product reviews tend to have high scores associated with them may indicate a propensity of reviewers to write only about products they are happy with, and/or may indicate the

[1] `http://www.tripadvisor.com/`
[2] `http://www.amazon.com/`

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.230

presence of many “fake” reviews (see [15] for a discussion) authored by people with vested interests. However, it is a fact that review datasets come in imbalanced form, and it is a fact that they are important[3]; as a consequence, automated systems that intend to mine them must cope with imbalance.

For standard (i.e., binary or multiclass) classification there is an entire strand of literature devoted to the consequences of imbalance. In classification applications, the main consequence of the imbalanced nature of a dataset is that, when a system is tested on it, an evaluation measure robust to imbalance must be used, i.e., a measure that does not reward “trivial classifiers” (see e.g., [16]). For a given (binary, ordinal, or other) classification problem a *trivial classifier* $\tilde{\Phi}_k$ may be defined as a classifier that assigns all documents to the same class $y_k$. Accordingly, the *trivial class for* $E$ (denoted $\tilde{y}$) may be defined as the class that minimizes the chosen error measure $E$ on the training set $Tr$ across all trivial classifiers, i.e.,

$$\tilde{y} = \arg\min_{y_k \in Y} E(\tilde{\Phi}_k, Tr)$$

Likewise, the *trivial-class classifier* for $E$ (denoted $\tilde{\Phi}$) may be defined as the trivial classifier that assigns all documents to the trivial class for $E$.

The need of penalizing trivial classifiers has long been acknowledged, e.g., in binary text classification [17], where it is often the case that the positive examples of a class are largely outnumbered by its negative examples (e.g., the Web pages about NuclearWasteDisposal are less than .001% of the total number of Web pages). A measure such as standard *error rate* (namely, the fraction of classified documents that are incorrectly classified) is not robust to this imbalance, since the *majority-class classifier* (i.e., the trivial classifier that assigns all documents to the majority class, which is the trivial class for error rate) would be deemed extremely “error-free”, probably more error-free than any genuinely engineered classifier.

For binary text classification the standard evaluation measure is $F_1$ [18], defined as the harmonic mean of *precision* ($\pi$) and *recall* ($\rho$), i.e.,

$$F_1 = \frac{2\pi\rho}{\pi + \rho} = \frac{2\frac{TP}{TP+FP}\frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} = \frac{2TP}{2TP + FP + FN}$$

where $TP$, $FP$, and $FN$ stand for the numbers of true positives, false positives, and false negatives resulting from classifying our set of documents. One of the reasons $F_1$ is standard is exactly because it is robust to this imbalance, since the majority-class classifier would obtain $F_1 = 0$ while the minority-class classifier would obtain an $F_1$ value equal to the frequency of the positive class, which is usually very low (less than .001, in the example above – note that for $F_1$, unlike for error rate, higher values are better)[4]. $F_1$ is thus preferred to error rate for evaluating binary classification [17].

The lack of robustness to imbalance on the part of an evaluation measure has two further negative consequences. The first is that, when a classifier depends on one or more parameters, optimizing these parameters on a validation set by using such a measure obviously returns parameter values that make the classifier behave very much like a trivial classifier. The second, related consequence is that, when a learning device is designed to internally optimize a given measure, or “loss” (as is the case, e.g., of SVMs or boosting-based learners), the resulting classifier may also resemble a trivial classifier.

## III. Common evaluation measures for OR

Are the standard evaluation measures for OR robust to imbalance? The most commonly used such measures are

1) *Mean Absolute Error* (here denoted $MAE^{\mu}$, and also called *ranking loss* – see e.g., [19]), as used e.g., in [20], [1], [21], [22], [23]. $MAE^{\mu}$ is defined as the average deviation of the predicted class from the true class, i.e.,

$$MAE^{\mu}(\hat{\Phi}, Te) = \frac{1}{|Te|} \sum_{x_i \in Te} |\hat{\Phi}(x_i) - \Phi(x_i)| \quad (1)$$

where $Te$ denotes the test set and the $n$ classes in $Y$ are assumed to be real numbers, so that $|\hat{\Phi}(x_i) - \Phi(x_i)|$ exactly quantifies the distance between the true and the predicted rank (the meaning of the $\mu$ superscript will be clarified later).

2) *Mean Squared Error* ($MSE^{\mu}$ – also called *Squared Error Loss*), as used e.g., in [14], defined as

$$MSE^{\mu}(\hat{\Phi}, Te) = \frac{1}{|Te|} \sum_{x_i \in Te} (\hat{\Phi}(x_i) - \Phi(x_i))^2 \quad (2)$$

A variant is *Root Mean Square Error*, as used e.g., in [21], which corresponds to the square root of $MSE^{\mu}$.

3) *Mean Zero-One Error* (more frequently known as *Error Rate*), as used e.g., in [20], [1], [21], [24], [12], [13], and simply defined as the fraction of incorrect predictions, i.e.,

$$MZOE^{\mu}(\hat{\Phi}, Te) = \frac{|\{x_i \in Te : \hat{\Phi}(x_i) \neq \Phi(x_i)\}|}{|Te|} \quad (3)$$

Unlike $MSE^{\mu}$ and $MAE^{\mu}$, $MZOE^{\mu}$ has the disadvantage that all errors are treated alike, and thus insufficiently penalizes algorithms that incur into blatant errors. $MSE^{\mu}$ penalizes blatant mistakes (e.g., misplacing an item into a rank faraway from the correct one) more than $MAE^{\mu}$, due

[3] A 2007 study (see http://www.comscore.com/press/release.asp?press=1928) found that between 73% and 87% “of review users (...) reported that the review had a significant influence on their purchase”.

[4] Here we assume that the positives are the minority and the negatives are the majority, which is usually the case in binary classification.

to the presence of squaring; as such, it has been argued (see e.g., [25]) that $MSE^{\mu}$ is more adequate for measuring systems that classify product reviews, since different reviewers might attribute identical reviews to different but neighbouring classes.

It is quite evident that none of these measures is robust to imbalance, since they are all based on a sum of the classification errors *across documents*. Since the majority-class classifier incurs in zero error for all the documents whose true class is the majority class, and since in an imbalanced dataset these documents are many, this trivial policy tends to be fairly "error-free".

To make this problem even worse, it is easy to show that for all these error measures the "trivial class" $\tilde{\Phi}_k$ need not be the *majority* class; in other words, there may exist trivial classifiers that are even more "error-free" than the majority-class classifier. For instance, in the TripAdvisor-15763 dataset mentioned above, assuming that the class distribution in the test set is the same as that in the training set, by assigning all test documents 4 stars we obtain lower $MAE^{\mu}$ than by assigning all of them 5 stars, which is the majority class. This is because 4 stars is only marginally less frequent than 5 stars, but in misclassifying all of the documents belonging to the lower classes (1 stars to 3 stars) as 4 stars we make a smaller mistake than in misclassifying them as 5 stars.

Little research has been performed in order to identify evaluation measures that overcome the shortcomings of measures (1)-(3). Gaudette and Japkovicz [25] acknowledge that these and other measures are somehow problematic but do not concretely propose alternatives. Waegeman et al. [26] instead propose an evaluation method based on ROC analysis. The problem with their method is that, like all methods based on ROC analysis, it is more apt to evaluate the ability of a classifier at correctly *ranking* the objects (i.e., at placing 5 stars reviews higher than 4 stars reviews) than to evaluate the ability of the classifier to classify an object into its true (or into a nearby) class. In other words, the ROC measure of [26] does not reward the ability of a learning device to correctly identify the thresholds $\tau_j$ that separate a class $y_j$ from its successor class $y_{j+1}$, for all $j = 1, \ldots, (n-1)$.

## IV. Making OR measures robust to imbalance

What can we do to make the three measures described in Section III more robust? The simple solution we propose is to transform them so that they are based on a sum of the classification errors *across classes*. This notion is inspired by the well-known distinction between the *microaveraged* and *macroaveraged* versions of $F_1$ (see e.g., [27]), where the former is obtained by averaging effectiveness across individual documents and the latter is instead obtained by first computing $F_1$ on a per-class basis and then averaging the results across the classes. According to this terminology, all the evaluation functions of Section III are microaveraged; we instead propose to use their macroaveraged analogues.

For instance, the macroaveraged version of $MAE^{\mu}$ (that we denote by $MAE^{M}$) is obtained by transforming (1) into

$$MAE^M(\hat{\Phi}, Te) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{|Te_j|}\sum_{x_i \in Te_j}|\hat{\Phi}(x_i) - \Phi(x_i)| \quad (4)$$

where $Te_j$ denotes the set of test documents whose true class is $y_j$ and the "M" superscript indicates macroaveraging (the "$\mu$" superscript we have used previously indicates instead microaveraging).

If consecutive ranks have always the same distance $d = |y_{j+1} - y_j|$, it is easy to show that the trivial class(es) for $MAE^M$ (and for $MSE^M$ and $RMSE^M$) are the middle classes, i.e. $y_{\lfloor \frac{n+1}{2} \rfloor}$ and $y_{\lceil \frac{n+1}{2} \rceil}$ (these coincide with the 3 stars class in the TripAdvisor-15763 and Amazon datasets discussed in this paper). For these classes the trivial classifier always obtains $MAE^M = \frac{n}{4}$ for even values of $n$ and $MAE^M = \frac{n^2-1}{4n}$ for odd values of $n$ (the trivial-class classifier thus obtains $MAE^M = 1.2$ in the both the TripAdvisor-15763 and Amazon datasets).

The effect of using $MAE^M$ on an imbalanced dataset (or any other dataset) is to make the trivial class for $MAE^M$ count as any other class, instead of proportionally to its frequency; assigning all test documents to the trivial class for $MAE^{\mu}$ produces zero error only for $\frac{|Te|}{n}$ test documents, which is not enough to guarantee low $MAE^M$.

A further interesting property of $MAE^M$ is that, on a perfectly balanced dataset, it coincides with $MAE^{\mu}$. In fact, given that on such a dataset its is true that $|Te_j| = \frac{|Te|}{n}$ for all $j = 1, \ldots, n$, we have

$$\begin{aligned} MAE^M(\hat{\Phi}, Te) &= \frac{1}{n}\sum_{j=1}^{n}\frac{1}{|Te_j|}\sum_{x_i \in Te_j}|\hat{\Phi}(x_i) - \Phi(x_i)| \\ &= \frac{1}{|Te|}\sum_{j=1}^{n}\sum_{x_i \in Te_j}|\hat{\Phi}(x_i) - \Phi(x_i)| \\ &= \frac{1}{|Te|}\sum_{x_i \in Te}|\hat{\Phi}(x_i) - \Phi(x_i)| \\ &= MAE^{\mu}(\hat{\Phi}, Te) \end{aligned}$$

Similar considerations hold for MSE and RMSE. We thus argue that macroaveraged versions of these measures should be the measures of choice in all OR contexts.

An example of the impact of using $MAE^M$ instead of $MAE^{\mu}$ comes from the following experiments (already described, although with different emphasis, in [11]). Table I reports the results of predicting the correct class (in a range from 1 star to 5 stars) of the product reviews in the TripAdvisor-15763 test set by means of an $\epsilon$-support vector regression ($\epsilon$-SVR) learning device [28] fed with standard bag-of-words representations. Recall that, as detailed in Section II, the class distribution of TripAdvisor-15763 is

Table I
$MAE^M$ AND $MAE^\mu$ RESULTS OBTAINED IN CLASSIFYING REVIEW DATA FROM ONE GLOBAL AND SOME SAMPLE LOCAL TRIPADVISOR-15763 DATASETS. "TRIVIAL" REFERS TO RESULTS OBTAINED BY THE TRIVIAL-CLASS CLASSIFIER FOR THE MEASURE INDICATED ($MAE^\mu$ OR $MAE^M$).

| Dataset | Method | $MAE^\mu$ | $MAE^M$ |
|---|---|---|---|
| Global | Trivial | 0.631 | 1.200 |
|  | $\epsilon$-SVR | 0.621 | 0.788 |
| Value | Trivial | 0.756 | 1.200 |
|  | $\epsilon$-SVR | 0.847 | 1.085 |
| Rooms | Trivial | 0.710 | 1.200 |
|  | $\epsilon$-SVR | 0.822 | 1.132 |
| Service | Trivial | 0.796 | 1.200 |
|  | $\epsilon$-SVR | 0.818 | 1.111 |

highly imbalanced. Since each review in the dataset has both a global score and other scores local to specific aspects (e.g., "Value", "Rooms", "Service", ...), each characterised by its own class distribution skew, experiments were actually run on both the global and other aspect-specific datasets.

Concerning the global dataset, we may see that, if using $MAE^\mu$ as a measure, $\epsilon$-SVR barely outperforms the trivial-class classifier for $MAE^\mu$ (.621 to .632, a mere +1.58% improvement); if using $MAE^M$, $\epsilon$-SVR outperforms the trivial-class classifier for $MAE^M$ by .788 to 1.200, a brisk +34.3% improvement. Concerning the "Value" dataset, if using $MAE^\mu$ our $\epsilon$-SVR is now even outperformed by the trivial-class classifier for $MAE^\mu$ (.847 to .756, a 12.0% deterioration); according to $MAE^M$, our $\epsilon$-SVR is instead better than the trivial-class classifier for $MAE^M$, although not by a wide margin (1.085 to 1.200, a 9.5% improvement). The "Rooms" and "Service" datasets behave similarly to "Value".

It is easy to guess that this problem might even be exacerbated on datasets, such as the above-mentioned Amazon dataset, in which the imbalance is even higher.

## V. CONCLUSIONS

We have proposed the use of macroaveraged versions of common measures such as mean absolute error or mean squared error, in order to cope with the imbalance problem in ordinal regression. These macroaveraged versions bring about robustness to imbalance and are equivalent to their standard microaveraged counterparts when the datasets are perfectly balanced. The adoption of these measures thus guarantees fair comparison among competing systems, and more correct optimization procedures for classifiers.

## ACKNOWLEDGMENT

This work was carried out in the context of the "Networked Peers for Business" (NeP4B) project, funded by the Italian Ministry of University and Research. We thank Nitin Jindal for clarifying the details of his experiments.

## REFERENCES

[1] W. Chu and S. S. Keerthi, "New approaches to support vector ordinal regression," in *Proceedings of the 22nd International Conference on Machine Learning (ICML'05)*, Bonn, DE, 2005, pp. 145–152.

[2] K. Crammer and Y. Singer, "Pranking with ranking," in *Advances in Neural Information Processing Systems*. Cambridge, US: The MIT Press, 2002, vol. 14, pp. 641–647.

[3] A. Shashua and A. Levin, "Ranking with large margin principle: Two approaches," in *Advances in Neural Information Processing Systems*. Cambridge, US: The MIT Press, 2003, vol. 15, pp. 937–944.

[4] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1/2, pp. 1–135, 2008.

[5] P. Beineke, T. Hastie, and S. Vaithyanathan, "The sentimental factor: Improving review classification via human-provided information," in *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL'04)*, Barcelona, ES, 2004, pp. 263—270.

[6] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, Prague, CZ, 2007, pp. 440–447.

[7] K. Dave, S. Lawrence, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," in *Proceedings of the 12th International Conference on the World Wide Web (WWW'03)*, Budapest, HU, 2003, pp. 519–528.

[8] A. Kennedy and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters," *Computational Intelligence*, vol. 22, no. 2, pp. 110–125, 2006.

[9] M. Koppel and J. Schler, "The importance of neutral examples for learning sentiment," *Computational Intelligence*, vol. 22, no. 2, pp. 100–109, 2006.

[10] A.-M. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP'05)*, Vancouver, CA, 2005, pp. 339–346.

[11] S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet rating of product reviews," in *Proceedings of the 31st European Conference on Information Retrieval (ECIR'09)*, Toulouse, FR, 2009, pp. 461–472.

[12] A. B. Goldberg and X. Zhu, "Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization," in *Proceedings of the HLT/NAACL Workshop on Graph-based Algorithms for Natural Language Processing*, New York, US, 2006, pp. 45–52.

[13] B. Pang and L. Lee, “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,” in *Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL’05)*, Ann Arbor, US, 2005, pp. 115–124.

[14] K. Shimada and T. Endo, “Seeing several stars: A rating inference task for a document containing several evaluation criteria,” in *Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD’08)*, Osaka, JP, 2008, pp. 1006–1014.

[15] N. Jindal and B. Liu, “Opinion spam and analysis,” in *Proceedings of the 1st International Conference on Web Search and Web Data Mining (WSDM’08)*, Palo Alto, US, 2008, pp. 219–229.

[16] N. V. Chawla, N. Japkowicz, and A. Kolcz, “Editorial: Special issue on learning from imbalanced data sets,” *ACM SIGKDD Explorations*, vol. 6, no. 1, pp. 1–6, 2004.

[17] Y. Yang, “An evaluation of statistical approaches to text categorization,” *Information Retrieval*, vol. 1, no. 1/2, pp. 69–90, 1999.

[18] D. D. Lewis, “Evaluating and optmizing autonomous text classification systems,” in *Proceedings of the 18th ACM International Conference on Research and Development in Information Retrieval (SIGIR’95)*, Seattle, US, 1995, pp. 246–254.

[19] B. Snyder and R. Barzilay, “Multiple aspect ranking using the good grief algorithm,” in *Proceedings of the Joint Conference of the North American Chapter of the Association for Computational Linguistics and Human Language Technology Conference (NAACL/HLT’07)*, Rochester, US, 2007, pp. 300–307.

[20] J. Basilico and T. Hofmann, “Unifying collaborative and content-based filtering,” in *Proceedings of the 21st International Conference on Machine Learning (ICML’04)*, Banff, CA, 2004, pp. 65–72.

[21] K. Dembczyński, W. Kotłowski, and R. Słowiński, “Ordinal classification with decision rules,” in *Proceedings of the ECML/PKDD’07 workshop on Mining Complex Data*, Warsaw, PL, 2007, pp. 169–181.

[22] R. K. Gopal and S. K. Meher, “Customer churn time prediction in mobile telecommunication industry using ordinal regression,” in *Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD’08)*, Osaka, JP, 2008, pp. 884—889.

[23] L. Li and H.-T. Lin, “Ordinal regression by extended binary classification,” in *Advances in Neural Information Processing Systems*. Cambridge, US: The MIT Press, 2007, vol. 19, pp. 865—872.

[24] E. Frank and M. Hall, “A simple approach to ordinal classification,” in *Proceedings of the 12th European Conference on Machine Learning (ECML’01)*, Freiburg, DE, 2001, pp. 145–156.

[25] L. Gaudette and N. Japkowicz, “Evaluation methods for ordinal classification,” in *Proceedings of the 22nd Canadian Conference on Artificial Intelligence*, Kelowna, CA, 2009, pp. 207–210.

[26] W. Waegeman, B. De Baets, and L. Boullart, “ROC analysis in ordinal regression learning,” *Pattern Recognition Letters*, vol. 29, no. 1, pp. 1–9, 2008.

[27] F. Sebastiani, “Machine learning in automated text categorization,” *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.

[28] V. Vapnik, *Statistical Learning Theory*. New York, US: Wiley, 1998.